Approaching text genre: How contextualized experience shapes task-specific performance

Petar Milin
University of Birmingham, United Kingdom

Filip Nenadić
University of Alberta, Canada

Michael Ramscar
Eberhad Karls University, Germany

Author Note
Petar Milin https://orcid.org/0000-0001-9708-7031
Filip Nenadić https://orcid.org/0000-0002-1450-5945
Michael Ramscar https://orcid.org/0000-0003-1680-1112

Correspondence concerning this article should be addressed to Petar Milin, University of Birmingham, School of Language, Cultures, Art History and Music, Ashley Building, Birmingham B15 2TT, United Kingdom. E-mail: p.milin@bham.ac.uk

Approaching text genre: How contextualized experience shapes task-specific performance

Abstract

In the present study, we sought to clarify how differences in contextualized experience influence the performance of participants engaged in genre decision-making. Using a simple learning algorithm, we ran a series of computational simulations to model the effects that context and cue competition have on the way readers of different backgrounds make genre decisions. Next, we used the results of those simulations as predictions for our behavioural genre decision experiment. Differences in test performance were strongly influenced by the factors that have long been known to influence learning: Cue competition and its embedding in a specific context jointly modulate what gets learned and that inevitably affects later performance. We discuss our findings in the context of learning and literary genres.

Key Words: genre, poetry, learning, experience, context, cue competition

"To know an object is to lead to it through a context which the world provides."
William James (1909), *The Meaning of Truth*

Many of the debates in the language sciences relate, at least in part, to the nature vs. nurture dichotomy (for a recent overview of this debate in language, see for example Ambridge & Lieven, 2011). Reading, however, seems to be an undisputedly nurtured, that is, *learned* behaviour . Reichle and Reingold (2013) convincingly described its stunning complexity, which we handle routinely and perform, literally, every day.

Fascinatingly, learning to read also entails being sensitive to and optimising for reading texts of different types. Consider how the way you approached this paper differs from the way you might approach a poem, a novel, or a sports page. Some of the specific reading differences associated with a scientific paper will stem from differences in text difficulty (Rayner, Chace, Slattery, & Ashby, 2006) and reading goals (Linderholm, 2006; Van den Broek, Lorch, Linderholm, & Gustafson, 2001). Yet, a text's type, commonly called genre or functional style, also exerts considerable influence on the way that text is read (cf., Hanauer, 1996, 2001). For example, differences in reading speed and recall are observed if participants are told that the *same* text is either a short story or a newspaper article (Zwaan, 1991, 1994).

Although initial genre categorization is usually implicit (Hanauer, 2001), it is also learned. Mastering this skill involves both implicit learning (as when a student learns that figures and tables are a good cue that a certain text is a scientific paper) and explicit teaching (as in when a teacher classifies a text as a scientific paper). This is exactly the reason why genre categorization represents an intriguing phenomenon that sits at the intersection of research on reading (e.g., Miall & Kuiken, 1998; Zwaan, 1991), studies of discourse (e.g., Bortolussi & Dixon, 2003; Miall, 2002), and even poeticity (e.g., Hoffstaedter, 1987; Peskin, 2007, 2010).

The genre categorization task that contrasts poetry and prose is the most often used task of this kind. In what follows, we present a concise summary of existing work on poeticity judgements. Then, we will focus on the role that learning plays in shaping these judgements, which will include details of how we learn to discriminate cues that are informative (or not) for deciding on a text's genre, and how this decision becomes possible through cue competition and is conditioned on the context in which it takes its place.

The Main Dimensions in Poeticity Decision-Making

One possible reason for using poetry vs. prose as the exemplary genre categorization problem is that poetry appears to be exceptionally distinctive across a range of characteristics, including the extent of exposure (Peskin, 2010) and the frequency of consumption (see, for example, a survey of literary reading in the USA by Bradshaw & Nichols, 2004). Yet, arguably, there are three factors that are of particular relevance for a poetry genre decision: (a) formal textual characteristics such as the presence of verses or stanzas; (b) text source – i.e., contextual information (e.g., explicit text labels, named source); and (c) participant experience. The effects of these factors have often been discussed in terms of the opposition between the *conventionalist* (source, expertise) and the *formalist* stances (objective text characteristics) in literary theory (an in-depth discussion can be found in Hanauer, 1996,

1998b). Currently, however, it seems that the fully formalist approaches to defining genres receive little support from literary scientists.

Differences between the formal or objective characteristics of texts, such as a particular layout or content (e.g., rhyme, alliteration), are reflected not only in poeticity ratings but also in reading performance more generally. For example, participants read poetry more slowly than encyclopaedia entries, and they also rate poetry as harder to understand, but they have better recall of the surface characteristics of poems (e.g., Hanauer, 1998a). Hanauer (1998b) used cued recall tests of words from texts to show that participants performed best when a text was presented in *high graphic form*, which includes stanzas and verses. Similarly, Peskin (2007, 2010) showed that transforming simple prose sentences into verses leads to changes in text interpretation and judging it as more *poetic*.

With regard to the text source, as means of conventionalization or contextualization, Hoffstaedter (1987) showed that participants rate texts as more poetic if they are told they come from a volume of poetry as opposed to being taken from a newspaper. It has been shown that the same factor also affects text comprehension even when other characteristics are kept identical (cf., Zwaan, 1993), as when the same text is once said to be literary and another time news (Zwaan, 1991, 1994).

Finally, experts and laypeople, being more or less experienced, exhibit differences in poeticity judgements (Peskin, 1998). These judgements can be further modulated by formal characteristics of the text – i.e., whether it is presented in verses (and stanzas) or not (Hanauer, 1996; Miall & Kuiken, 1998), as experts rate texts with typical graphic (stanzas and verses) and phonetic (rhyme and alliteration) characteristics significantly more poetic than laypeople do.

Taken together, factors such as the text form and the text source as well as the reader's experience crucially affect genre categorization and, more broadly speaking, evaluation across the whole range of relevant text characteristics (like, for example, text difficulty, viz. Rayner et al., 2006). Interestingly, however, as with reading itself, these three factors ought to be learned too: They are useful, salient cues for making a well-informed decision about a text's genre. How does this learning take place, and how can learning theory, more generally, and certain learning phenomena, specifically, contribute to better understanding this important, early optimizer of reading behaviour and of uniquely human aesthetic experience?

Learning, Performance and Poeticity Decision-Making

For the purpose of the present study, we will define learning as the process of acquiring information about relations between events and/or objects in a particular environment (see Rescorla, 1988). Learning is, thus, (a) a *contextualized experience* with or about (b) *cues competing for predictive significance* about events, objects and the environment itself.

Importantly, experience must accrue to allow for learning to show its effectiveness (Gallistel & Gibbon, 2000; Kimmel, 1959; Rescorla, 1970), and it will inevitably take place in a specific context (which is, in the learning framework, also referred as background or ambient; cf., Bouton, 1993; Kruschke & Hullinger, 2010). This means that there are two important components of experience that we need to distinguish: its *quantity* and *quality*, where the former refers to the amount of learning underlying a certain behaviour, and the latter relates to the context in which that learning is situated. Ultimately, context and experience – the

qualitative and quantitative aspects of the learning process – both exert a similar influence, which is to (a) inhibit the learning of new associations while simultaneously (b) facilitating performance on tasks which invoke relations that have already been learned (see Hull, 1943; Konorski, 1967; Rescorla, 1970; Rescorla, Durlach, & Grau, 1985; Spence, 1956; Tolman, 1949).

Given this 'conflicting' inhibitory-facilitatory disposition and, in particular, the highly contextualized nature of cue competition, learning can have *both positive and negative effects*. It may appear as failing us in novel or altered situations, leading to erroneous predictions. This is because learning, being "always in effect" (cf., Spreat & Spreat, 1982, p. 593), continuously (re-)calibrates the predictions against the error to enhance adaptation. Consequently, prediction can be over-specialized, constrained to and effective in very specific circumstances only.

Although it is not always described as such, evidence of the effects of the negative consequences of cue competition in learning on later performance abounds elsewhere in the cognitive science literature: Expert chess players are sometimes worse at remembering impossible chess board configurations than novices (Chase & Simon, 1973); radiologists perform at a chance level once X-ray images are inverted (Bilalić, Grottenthaler, Nägele, & Lindig, 2014); once important perceptual cues are occluded, expert tennis players lose their ability to predict the landing position of balls (Abernethy & Russell, 1987); and despite their spatial expertise, London taxi drivers actually perform worse at novel object location paired associate learning tasks – and those tasks alone – as compared to controls (Woollett & Maguire, 2009, 2011).

At least some of these intriguing and counter-intuitive empirical phenomena of 'failing expertise' seem to be related to the *context switch effect* in learning (Rosas, Todd, & Bouton, 2013), which emerges "when contexts have been given predictive or informative value," and "primarily when *experience* encourages attention to [them]" (Rosas et al., 2013, p. 241, emphasis added). The effects are particularly pronounced when the prediction error increases, which is why Bouton (1997) as well as Rosas, Aguilera, Álvarez, and Abad (2006) used *ambiguity* to explain that, experientially, high uncertainty (i.e., ambiguity) leads subjects to make use of the *context as weighting element* that might minimise prediction errors.

Can literary expertise lead to similar failures in genre categorization? And, more broadly, can experience and learned cues for genre categorization trick us into making the wrong decision which would ultimately lead to sub-optimal reading? We hypothesize that with more experience (quantity) across potentially more diverse contexts (quality), 'genre specialists' will in fact make more overlapping or blurred categorizations, particularly when, like in most of the studies of experts' omissions, the most salient cues are removed and/or crucial information is otherwise altered or distorted. Indirect but convincing support for our predictions can be found in the literature. The effect of context emerged in the study of Hoffstaedter (1987), who showed that participants rated texts to be more poetic simply because they were told that those texts were taken from a volume of poetry as opposed to being newspaper articles (compare this with the results of Zwaan's studies in 1991 and 1994). Similarly, the effect of experience is demonstrated in the work of Hanauer (1996), demonstrating specific differences between poetry experts and laypeople.

Present Study

Our study starts from the assumption that readers (individually and as a community more broadly) form categories (genres) of texts to optimize reading. Similar to learning many other kinds of socially sophisticated cues, learning cues that reliably predict genre categories depends on the quantity and the quality of the experience with such cues and the outcomes they are associated with in a given environment. As the quantity of experience accrues, what is learned becomes more stable, yet it seldom (if ever) gets fixed. The same general principle applies to learning genre categories, too: The more experience with texts of the same basic types, the more solid the genres appear to be. Even so, if we purposefully diversify or otherwise alter the contexts in which these texts are experienced, we may observe a change in behaviour, e.g., genre categorization becomes less clear-cut. This essentially shows how the context *switches* or *weights* what is learned, for the purpose of adaptation to new conditions.

It seems likely that a poetry expert will not only have more but also more diverse experience of reading poems. For most readers, schooling drives the learning of which forms are cues to which literary genres, and later (higher education) experiences may (or may not) serve to revise these early-acquired models (cf., Peskin, 2010). Accordingly, if we were to compare the experiences of college-age literature students and STEM students, we might expect to find a fair degree of congruence in their experiences up to the end of secondary/high school, and a concomitant divergence in their experience thereafter, as they begin to specialize in their respective disciplines. In addition to the fact that they will accumulate more experience, we may also expect that literature students will encounter poetry in more diverse contexts. This will make their genre decision making less 'crisp' and distinguishable. We will argue that this is not at all a failure but a mere consequence of intensive and diverse learning.

To simulate the aforementioned assumptions, we will make use of a basic algorithmic insight into the learning processes (Rescorla & Wagner, 1972; Widrow & Hoff, 1960) which is the foundation for a computational model of discrimination learning (Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011; Milin, Feldman, Ramscar, Hendrix, & Baayen, 2017; Milin, Madabushi, Croucher, & Divjak, 2020; Ramscar & Yarlett, 2007; Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010; Rescorla, 1988; Rescorla & Wagner, 1972). We will systematically manipulate the amount (quantity) of experience and the diversity of contexts (quality) to observe cue competition and to derive assumptions about the effect that learning will have on genre decision-making. However, while manipulating the amount of experience is rather straightforward, there are many approaches in the learning literature to formalizing context (cf., Landeira-Fernandez, 1996; Myers & Gluck, 1994; Nadel & Willner, 1980; Urcelay & Miller, 2014). We have opted for a maximally simple operationalization: (a) the context will be implemented as any other cue (following Tomie, 1981, who showed that this provides a surprisingly accurate functional characterization for most modelling purposes); (b) its specificity lies in the fact that it serves as a longer-term or static rather than as a punctuate cue (following Rescorla et al., 1985); finally, (c) it serves as weighting element for other cues that compete for 'a chance' to minimise the prediction error (following Rosas et al., 2006; see Kruschke & Hullinger, 2010, who have explicitly argued that context is directly predictive only of other cues which can be directly predictive of an outcome).

Finally, the results of our simulations will yield a (formal) prediction for the experimental part of this study investigating the effects that amount of experience and diversity of learning context have on genre decisions in a categorization task. In the behavioural experiment, we represent each of the three important aspects of genre decision using a single prominent factor with a confirmed effect in previous studies: We manipulate text layout, text source, and test participants with different backgrounds.

Empirical Study

To test the importance of context and prior experience on cue competition and, ultimately, on cognitive performance in a genre decision task, we designed a two-fold task: We first ran computational simulations to generate specific predictions that were then tested experimentally. The simulations represented an idealised learning situation for two groups of learners that would differ in both the quantity (i.e., amount) and quality (i.e., diversity of learning contexts) of their experience with two different text genres. We postulated a degree of congruence in the groups' early experiences, and then a divergence later on. The results of the simulations were used as predictions which were compared with the results obtained in an experimental study, where we tested the performance of college age literature students and STEM students on a genre decision task.

Computational Simulation

To model differences in experience, we made use of perhaps the simplest algorithm for modelling cue competition in learning (RW: Rescorla & Wagner, 1972; or WH: Widrow & Hoff, 1960). As well as being biologically plausible (Chen, Haykin, Eggermont, & Becker, 2008), the learning principles it embodies have been shown to be superior in an evolutionary sense to many other, more complex mechanisms of learning (see Trimmer, McNamara, Houston, & Marshall, 2012). While this rule does not exhaust the human capacity for learning, it can be safely assumed to set a minimum threshold for it.

This algorithm represents the simplest learning network – a special case for many other, more complex learning networks. For example, Kalman Filtering is its Bayesian extension (Kakade & Dayan, 2002; Kalman, 1960; Sutton, 1992), while Temporal Difference Learning generalizes to model continuous updating as well as within-trial relationships (Dayan & Niv, 2008; Niv, 2009; Sutton & Barto, 1990). RW/WH also captures principles of learning and generalization that have been successfully applied to the domain of memory and retrieval (Hintzman, 1984, 1988).

From a functional perspective, the RW/WH model implements a simple yet highly effective procedure for enforcing cue competitions (a basic principle of discrimination learning) making it an obvious (straightforward and parsimonious) choice for computational modelling, given our theoretical grounding and empirical goals. By restraining from manipulating its parameters, but by carefully defining and feeding its input and output and, crucially, the amount and context of experience, this simple algorithm will yield predictions that can be falsified and corroborated with experimental evidence.

A Simplified Widrow-Hoff / Rescorla-Wagner Model

The model's equations describe a discrimination learning algorithm, defining the way that the evidential weights of a *system of cues* (e.g., the appearance of a text, or a book, a classroom, etc.) in relation to a *set of outcomes* that represent experiential and behavioural contrasts (e.g., judging a text to be a poem) change with experience. The relation is recurrent, in discretized time. Let $PRESENT(C_i, t)$ denote the presence of a cue C_i in a given learning event taking place at time t, and let $PRESENT(O_J, t)$ denote the presence of an outcome O_j in event at time t. Let then w_{ij}^t denote the weight from C_i to O_j at time t. The weight at the subsequent timestep w_{ij}^{t+1} is consequently defined as

$$w_{ij}^{t+1} = w_{ij}^t + \Delta w_{ij}^t \tag{1}$$

where the change in weight Δw_{ij}^t is specified as

$$\Delta w_{ij}^t = \begin{cases} 0 & \text{if } PRESENT(C_i, t) \text{ is } \textbf{FALSE} \\ \gamma\left(\lambda_j - \sum_{PRESENT(C_k,t)} w_{kj}\right) & \text{if } PRESENT(C_i, t) \text{ is } \textbf{TRUE} \ \& \ PRESENT(O_j, t) \text{ is } \textbf{TRUE} \\ \gamma\left(0 - \sum_{PRESENT(C_k,t)} w_{kj}\right) & \text{if } PRESENT(C_i, t) \text{ is } \textbf{TRUE} \ \& \ PRESENT(O_j, t) \text{ is } \textbf{FALSE} \end{cases} \tag{2}$$

In (2) the parameter λ_j denotes the maximum amount of learning that a given outcome can receive – its 'learnability.' For the present simulation, we set its values to $\lambda_j = 1.0$, which means that we simply normalize the end-result of learning and make all outcomes comparable as to how well they have been learned (where, ideally and in the limit, some of them may be learned perfectly – i.e., $w_{ij} = 1.0$). Parameter γ gives the *learning rate*, which we set to $\gamma = 0.01$, a desirably small value that guarantees learning to be incremental (for further details see Blough, 1975; Enquist, Lind, & Ghirlanda, 2016; Rescorla & Wagner, 1972).

Given the change in weight Δw_{ij}^t, the RW/WH equations state that when a cue is not present, the weights on its connections to the outcomes remain unchanged. When a cue is present, and a given outcome is present, the weight on the connection from the cue to that outcome is strengthened as a function of the decrease in uncertainty associated with the outcome. Conversely, when a cue is present without an outcome, the connection weight is weakened according to the same function, as the uncertainty about the outcome increases. In simple terms, this means that if a cue and outcome co-occur, the linking weight is increased (*learned*), but if a cue occurs without the outcome, this weight is decreased (*unlearned*). This process serves to minimize the prediction errors generated by the set of cues as a whole. Accordingly, the total number of relevant cues has a large impact on learning. As the number of cues increases, the amount by which a weight is increased is reduced. By contrast, the more cues are present, the greater this decrease in a given weight.

Simulation Setup and Results

The aim of the computational simulation was to capture the typical differences that can be expected to occur between the life-experiences of Literature students and STEM students, and to examine how these differences might affect the students' genre decisions and the reading that follows this decision. The differences in students' exposure to text-types were operationalized as follows: Two groups of students (Literature vs. STEM) were assumed

to have a different number of learning trials (i.e., the *amount of experience*): $TOTAL_{LIT} = 6\,000$ and $TOTAL_{STEM} = 4\,000$ respectively. Two input cues (*text1* and *text2*) were specified, in order to represent differences in the form of prose and poem texts (e.g., whether a text is organized into verses and stanzas or not), which are usually experienced from school age on, along with two possible text classification outcomes, one which we shall simply refer to as *prose* and *poem*. In addition to that, *context of experience* was defined as either *anthology* or *reading book* (*r.book*), which corresponds with text source. The latter was set as more ambiguous with respect to the associated genre. At each time step, one of the two cues (*text1* or *text2*) and one of the two contexts (*anthology* or *r.book*) was linked with one of the two outcomes (e.g., *r.book* + *text2* → *prose*).

Experience was distributed over three broad age-related phases: *childhood* (1-7), *school-age* (8-18), *college/studies* (19-22), and learning events were randomized within these three blocks. Table 1 summarizes the learning setups for both study groups. Crucially, there is uncertainty in the *text2* cue in the *r.book* context, and importantly, given the differences in their later experience of texts, this uncertainty is more pronounced (crucially, not extremely more) for Literature students than for STEM students (0.3 : 0.7 vs. 0.1 : 0.9). This uncertainty was, however, kept exactly the same for both groups in the first two phases – *childhood* and *school-age* (0.15 : 0.85). This effectively means that the average experience only started to diverge when their studies specialized: that is, when they entered higher education.

Table 1. Summary of learning events set-up for the first simulation of experiences of the two groups of students (in relative frequencies).

Study group	**Context**	**Cues**	**Outcomes**	
			poem	*prose*
STEM	*anthology*	*text1*	1.0	0.0
		text2	0.0	0.0
	r.book	*text1*	1.0	0
		text2	0.1	0.9
Literature	*anthology*	*text1*	1.0	0.0
		text2	0.0	0.0
	r.book	*text1*	1.0	0
		text2	0.3	0.7

Figure 1 shows the development of the discrimination (i.e., associative) weights between the cues and the outcomes in the two models over trials. Note the tendency for *text1* to cue *poem* and for *text2* to cue *prose*. For the STEM student model, the two contexts provide further support for this tendency, with *anthology* as somewhat more 'poetic' and *r.book* as more 'prosaic.' The weighting support from the two contexts (*anthology* and *r.book*) for the two form cues (*text1* and *text2*) is presented in Figure 2.

For STEM students, *text1* predicts *poem* and *text2* predicts *non-poem* (given the negative weight values). The context of *anthology* gives heavier weighting (both, more positive and more negative) to such predictions than *r.book*, but overall, for this group of students, context does not make predictions any more uncertain. For the Literature student model, however, *text1* provides high certainty in predicting a genre – *poetry*, irrespective of

a particular context. In fact, *r.book* makes the predictions even more pronounced. Yet, for this group of students the development of the discrimination weights for both *r.book* and *text2* have not stabilized by the end of the simulation period (compare the solid ochre and dotted magenta lines in the upper and lower panels of Figure 1, representing the models for STEM and Literature students). The uncertainty is also reflected in the support for predicting a *poem* where we observe less systematic weighting of specific form cues by the two contexts: Experience has led to the unlearning of any strong connection from *text2* to a specific genre classification in the Literature student model (see Figure 2a).

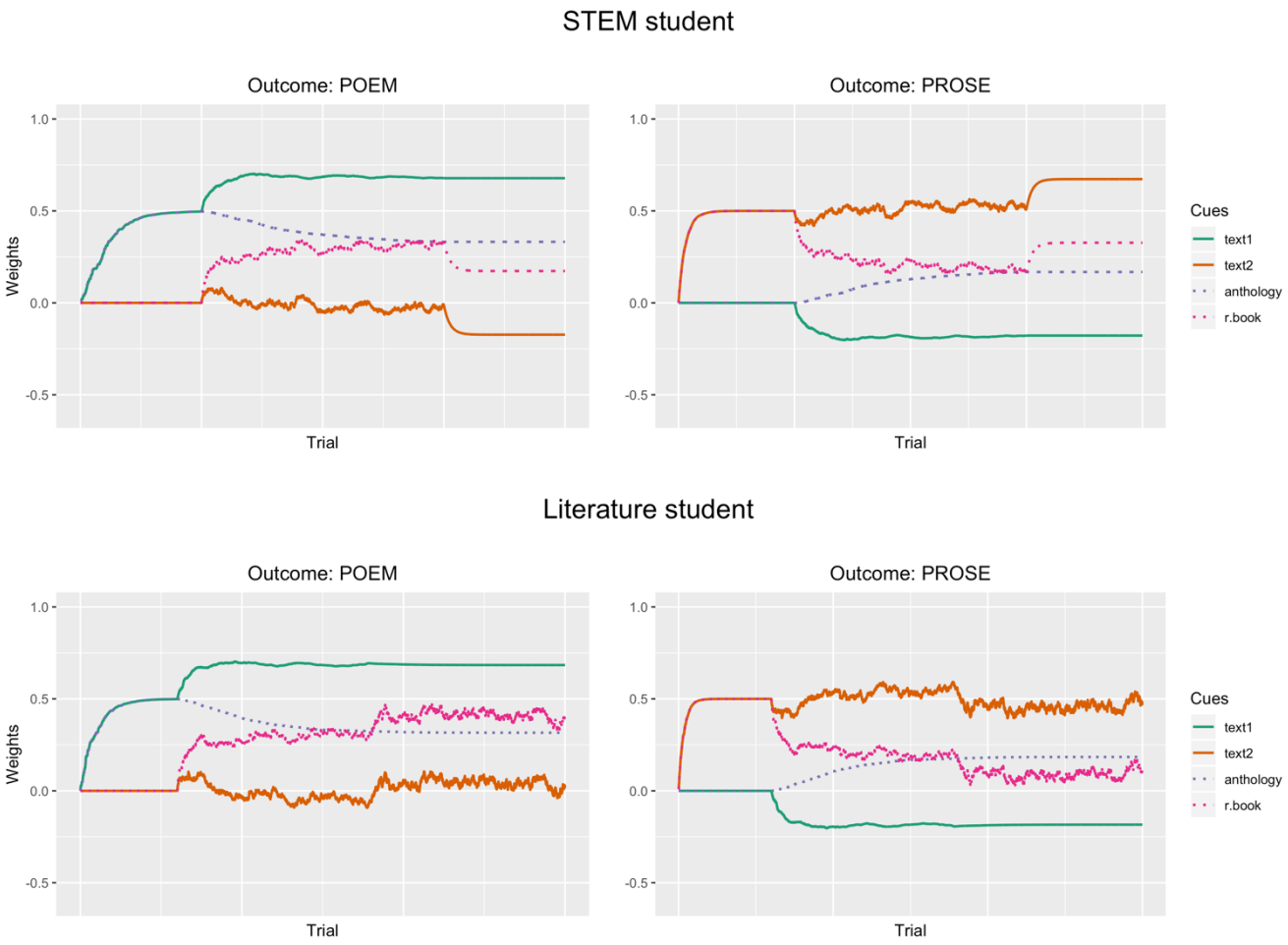


Figure 1. Development of discrimination weights from two form cues and two contexts to two outcomes, in 4000 and 6000 learning epochs for STEM and Literature students, respectively.

Some further manipulations of the learning set-up for the two student models were conducted to help illustrate the effects of the amount and context of the learning experience, i.e., its quantity and quality. First, we changed the learning session and left in only the two later, age-related phases (i.e., *school-age* and *studies*), which effectively reduced the amount of experience (from 6 000 and 4 000 to 4 800 and 3 000 learning trials for Literature and STEM students respectively). The results of this second simulation are presented in Figure 2b. Most weights remained unaffected by the reduction in the amount of experience, except for some miniscule changes (e.g., *text1* as weighted by the *anthology* context becomes equally

predictive of *poem* for both groups of students) which can be ignored here. The second simulation illustrates the point that the *amount* (quantity) of experience is not a simple (or single) influencing factor for *cue competition*: Some experience is indeed necessary for *systematicity* in cue competition to emerge; yet, systematicity itself is sufficient for the discrimination (association) weights to stabilize and gain their true predictive value.

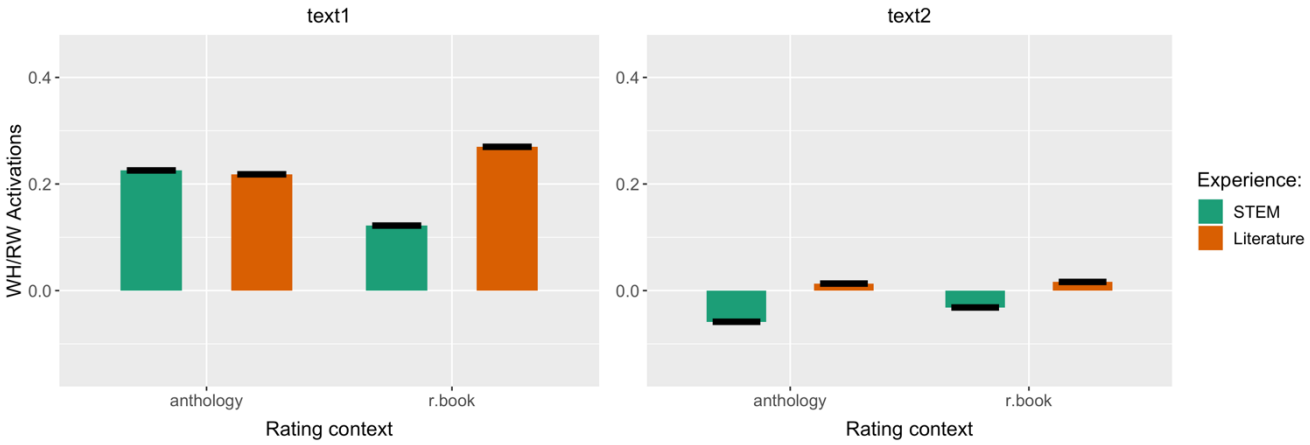


Figure 2a. Context-weighted support for form cues (text1 and text2) for the prediction of poetry. Weighting is obtained by multiplying the form cues with the respective contextual cues (anthology and r.book). The black horizontal bars are values of support obtained in this benchmark simulation, for comparison with the results from other simulations.

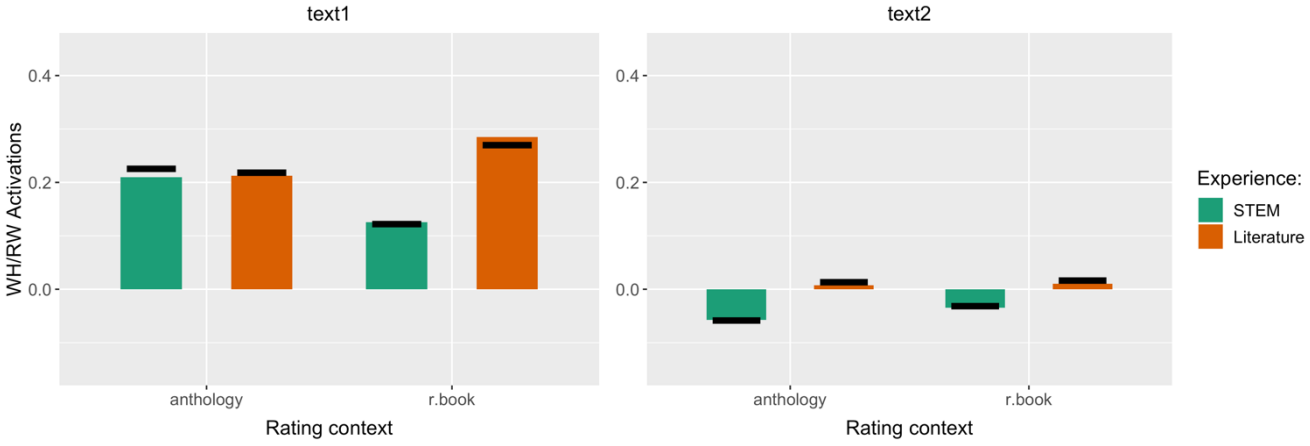


Figure 2b. Context-weighted support for form cues (text1 and text2) for the prediction of poetry genre, after reducing the amount of experience. Weighting is obtained by multiplying the form cues with the respective contextual cues (anthology and r.book). The black horizontal bars show the values from the benchmark simulation presented in Figure 2a.

Taken together, the results from the three learning simulations yield a (formal) prediction that we will now test experimentally: Students of Literature will be less certain that simple visual cues are sufficient for making a genre decision about a text, and consequently,

when faced with the task of classifying an unknown text in a specific context, they will be far less certain about *text-X → poem* classifications than STEM students.

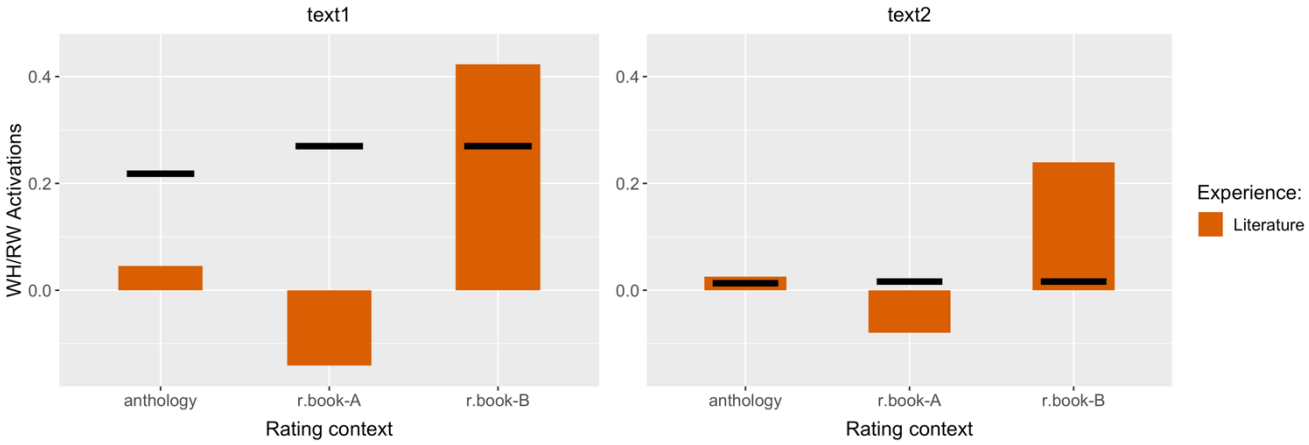


Figure 2c. Context-weighted support for form cues (text1 and text2) for the prediction of poetry, after diversifying the quality (context) of learning. Weighting is obtained by multiplying the form cues with respective contextual cues (anthology, r.book-A, and r.book-B). The black horizontal bars show the values from the benchmark simulation presented in Figure 2a.

The three simulations yield straightforward predictions that will be tested against experimental data. First, the amount (quantity) of experience is essential to establishing the baseline systematicity between the events in the environment. Second, the contextualisation (quality) of experience will alter the cue competition itself, by clearing or blurring the predictability in a given situation. Thus, the quantity appears to be important for the *validity* of the learning-based predictions while the quality influences the *reliability* of these predictions. In other words, quantity is important for *accuracy* and quality for *precision*, which means that both are crucial for performance on any given task.

In the context of the present study, we can assume that both groups of students (Literature and STEM) will have sufficient experience with different texts and their genres. This means that students have already established baseline accuracy in their genre decisions. Literature students will, however, have a significantly more diverse experience which will, in turn, make their genre decisions more diverse. In other words, their decisions will be more varied and this variation should be even more pronounced if cues that are otherwise particularly salient sources of relevant information for a fully informed and precise decision are absent (e.g., the content of a given text). Inasmuch as this prediction might initially appear counterintuitive, it is only so if we trap ourselves by thinking about decision-making in absolute terms. The point here is that Literature students have experienced “genre” as complex and nuanced, making simple, binary choices undesirable.

Genre Decision Experiment

Our model of genre learning makes clear predictions about the importance of context and the amount of prior experience on performance. Other textual characteristics were carefully controlled for by blocking such cues and keeping them constant to the extent possible.

Materials and methods

Participants. 64 students of the University of Novi Sad (Serbia), aged 22 to 27 were tested. Half of the participants were postgraduate students of Literature and, consequently, had more extensive prior experience in reading and classifying literary texts. The other half of students were age-related peers from STEM subjects, who confirmed that they were neither poetry or literature enthusiasts nor had been required to read poetry outside of elementary and high school.

Study design. The experiment followed a 4x4x2 factorial design. Text Form had four levels: *stanzas* (stanzas, verses, capital letter at the beginning of each verse); *verses* (one stanza, verses, capital letter at the beginning of each sentence); *left* aligned prose text; *justified* text alignment. Participants were told what the text Source was, in the form of a short introduction: *anthology* (anthological poetry); *r.book* (secondary school reading book); *article* (newspaper article); *text* (generic text with no specifications given). Finally, two prior experience groups were determined by our participants' respective degree programs: *Literature* vs. *STEM* students (see above). Three dependent variables were measured:

1. Binary responses (yes/no) to a choice question in which participants decided whether the text is a poem or not.
2. Rating the degree to which they felt a text was a poem ranging from -3 (not at all a poem) to 3 (definitely a poem), including 0 (undecided).
3. Confidence ratings for the previous two responses, ranging from 1 (not at all confident) to 5 (completely confident).

The questions were designed to be redundant to allow us to establish consistency in participants' answers. For example, a participant could respond that a stimulus was a poem in Q1 (answering yes), and yet rate the degree to which they felt it was a poem rather low in Q2 (rating between -3 and -1). This enabled us to detect inconsistencies in answers (e.g., "yes" on Q1 and a low rating on Q2 or vice versa). Inconsistencies were found only in 0.01% of all cases, indicating that our participants were confident about and consistent in their ratings. Since answers to Q1 and Q2 were consistent, we can focus our analysis on the second question that provided participants' ratings of a text being 'poetic' (-3 to 3).

Stimuli. Stimuli were original poems or short stories written in Serbian or translated into Serbian that had been made unreadable by blurring the words. Appendix A provides an example of an experimental item. Texts were blurred in **LaTeX**, using the **\lastbox** and **\duplines** functions that produced a multi-layered text with minor offsets, which then appeared as blurred and unfit for reading. Such illegible texts were used to control for 'meaningful reading' confounds; in effect, we removed phonetic, morphologic, syntactic and semantic characteristics (i.e., most of what is related to the actual words, phrases, and meanings). Again, as explained above, this manipulation was implemented to annul the cue

that is, arguably, the most salient in making a genre decision, while promoting the experience, context and competition of the targeted genre cues. Twenty such texts were generated for each of the four types of text form, for a total of 80 texts. Four experimental lists were formed by grouping five texts of each of four text forms, for a total of 20 texts in a single list.

Procedure. The task in all four experimental blocks was the same: An illegible, blurred text was presented; participants decided whether it was a poem, and then rated the extent to which it appeared to be a poem to them (from -3 to 3) as well as their confidence in the response (from 1 to 5). The experiment was built using **OpenSesame** experimental software (Mathôt, Schreij, & Theeuwes, 2012). It was conducted in individual sessions which lasted no longer than 15 minutes each. The experiment was divided into four blocks, one for each of the four experimental lists. The blocks and the order in which the text forms were presented within blocks was counterbalanced across participants. At the start of each block, participants received concise, standardized information about the supposed origin of the text stimuli, i.e., the text source (anthology, reading book, newspaper, text). There were twenty trials within each block, containing five examples of each of the four types of text form. Trials were randomized, and responses were not time limited. Once finished with an experimental block, the participant proceeded to the next, receiving new information about the text source. This procedure was repeated until all four lists were rated.

Results

On average, participants saw the stimulus, decided whether it was a poem, rated the extent to which it was a poem and, finally, assessed their confidence in their answer in under five seconds. The grand mean for the yes/no question was 2817ms, and the other two responses were delivered even faster, at 909ms for the 7-unit ratings, and 670ms for the confidence ratings. These fast responses were consistent with the high degree of confidence participants exhibited in their ratings on the third question, with the median confidence rating being 4 out of 5.

Data provided in answer to question 2 (“To what extent is the text a poem?”) was analysed using Generalized Additive Mixed Modelling (GAMM) in the statistical software environment **R** (R Core Team, 2017), utilizing packages **mgcv** (Wood, 2007) and **itsadug** (Van Rij, Wieling, Baayen, & van Rijn, 2015). We applied GAMM with the **ocat** link-function that implements regression for ordered categorical data. In particular, a latent variable follows a logistic distribution and its probability reflects the ordered categorical variable being of the corresponding category (for details consult Wood, Pya, & Säfken, 2016). To test our main prediction that the responses of literary students are more diverse we used Wald’s test for contrasts.[1] We additionally applied Bonferroni’s correction for multiple comparisons, to remain conservative in the statistical sense.

The model was, *de facto*, an ANOVA, specifying our 4x4x2 factorial design: (1) text Form (stanzas, verses, left aligned, justified); (2) text Source (anthology, reading book, article, generic text); (3) prior Experience (Literature vs. STEM students). Additionally, the random

[1] The package **itsadug** can do Wald’s comparisons only between two specified levels. For this reason we developed our own function which allowed for much greater flexibility in comparisons, and for comparisons between combinations of factor levels more specifically.

effect part contained: (1) intercept adjustments for experimental items; participants' factor-level adjustments (i.e., slope adjustments) for both (2) the Form and (3) the Source. This elaborate random effect structure ensures that only the most robust fixed effects emerge as significant. Finally, the fixed effects contained two critical two-way interactions for the targeted comparisons and, ultimately, for hypothesis testing: Experience by Form, and Experience by Source. The model formula was specified as follows:

$$
\begin{aligned}
Rating \sim\ & \\
& Form \times Experience + \\
& Source \times Experience + \\
& s(Participant, Form, bs =' re') + \\
& s(Participant, Source, bs =' re') + \\
& s(Item, bs =' re')
\end{aligned}
$$

All model terms – the main fixed effects and interactions as well as the random effects adjustments for intercepts and factor levels – came out as significant. Following a forward model fitting strategy and using the Akaike information criterion (AIC) and its derivative the Evidence Ratio ($exp(abs(AIC_1 - AIC_2)/2)$) as measures of goodness-of-fit for selecting the best models, we confirmed that the final model was by far superior to any alternative. For the fixed effect part, specifically, the model above had a better AIC and Evidence Ratio than two alternative models containing only one of the two interactions:

Full model	**Alternative, simpler model**	**AIC Diff.***	**Evid. Ratio***
$Form \times Experience +$ $Source \times Experience +$	$Form +$ $Source \times Experience +$	66.46	2.70e+14
	$Form \times Experience +$ $Source +$	30.68	4,593,320

* In favour of the full model.

For the random effect part, the final model also showed a much better fit than alternative models with a simpler random effects structure:

Full model	**Alternative, simpler model**	**AIC Diff.***	**Evid. Ratio***
$s\left(\begin{matrix}Participant, Form,\\ bs =' re'\end{matrix}\right) +$ $s\left(\begin{matrix}Participant, Source,\\ bs =' re'\end{matrix}\right) +$	$s\left(\begin{matrix}Participant,\\ bs =' re'\end{matrix}\right) +$ $s\left(\begin{matrix}Participant, Form,\\ bs =' re'\end{matrix}\right) +$	537.34	4.80e+116
	$s\left(\begin{matrix}Participant,\\ bs =' re'\end{matrix}\right) +$ $s\left(\begin{matrix}Participant, Source,\\ bs =' re'\end{matrix}\right) +$	981.64	1.44e+213

* In favour of the full model.

The summary table with the results of the GAMM fitted to participants' ratings (Q2) is provided in Appendix C (Table C.1). To test our main hypothesis that a more diverse experience (in quality) leads to decisions that are more diverse too, we compared whether the differences in ratings across different Forms as well as different Sources were less

pronounced for the Literature students than for the STEM students, using targeted Wald testing for contrasts with Bonferroni correction for multiple comparisons. First, we ordered the effects of Form (stanzas, verses, left aligned, justified) and Source (anthology, reading book, article, text), separately by Experience (Literature vs. STEM students). Then, we applied Helmert's contrast coding which compares "each [factor] level with the average of the 'preceding' ones" (cf., Fox, 2002, p. 128). The results of all relevant Wald contrasts are summarized in Table 2.

Table 2. Wald's test for contrasts with Bonferroni p-value corrections, using Helmert's contrast coding scheme.

Contrast	**STEM students**	**Literature students**
Form:		
[justified] vs. [left al.]	$\chi^2 = 3.021;$ $p_{Bonf.} = 0.9866$	$\chi^2 = 0.253;$ $p_{Bonf.} = 1.0$
[justified, left al.] vs. [verses]	$\chi^2 = 105.140;$ $p_{Bonf.} < 0.0001$	$\chi^2 = 12.246;$ $p_{Bonf.} = 0.0056$
[justified, left al., verses] vs. [stanzas]	$\chi^2 = 208.366;$ $p_{Bonf.} < 0.0001$	$\chi^2 = 7.115;$ $p_{Bonf.} = 0.0917$
Source (Context):		
[text] vs. [article]	$\chi^2 = 10.785;$ $p_{Bonf.} = 0.0122$	$\chi^2 = 0.889;$ $p_{Bonf.} = 1.0$
[text, article] vs. [r.book]	$\chi^2 = 8.605;$ $p_{Bonf.} = 0.0402$	$\chi^2 = 3.880;$ $p_{Bonf.} = 0.5863$
[text, article vs. r.book] vs. [anthology]	$\chi^2 = 9.498;$ $p_{Bonf.} = 0.0247$	$\chi^2 = 14.922;$ $p_{Bonf.} = 0.0013$

The results reveal substantial differences in contrasts between the two groups of students: STEM students show a very clear separation across all contrasts, while Literature students show a more muddled picture, with considerable overlap (i.e., greater variation) in their answers. More precisely, STEM students showed no difference only in one case (between Form of justified vs. left aligned), while Literature students showed yes-difference only in two cases (between Form of verses vs. justified and left aligned, and between Source of anthology vs. r.book, article, and text). This change, to put it differently, from *mostly significant differences* for STEM students to *mostly non-significant differences* for Literary students is what makes both two-way interactions (Experience by Form and Experience by Source) highly significant in the GAMM model.

Finally, the tight structure of random effects containing participants' factor level adjustments for Form and Source enables us to test our hypothesis in a second way. Random effect adjustments are not parameter estimates but estimates of variance (or deviation) around a given fixed effect. Since each participant has their specific adjustments for different Forms and different Sources, it is possible to test whether these adjustments are greater for the group of Literature students than for STEM students. We found that, for both Form- and Source-related adjustments, Literature students showed consistently greater variance (Form: 1.279 vs. 0.804; Source: 0.626 vs. 0.483). The difference in Form-related variance was

statistically significant ($F_{(127,127)} = 1.591;\ p < 0.01$), while the difference in Source-related variance was not ($F_{(127,127)} = 1.298;\ p = 0.14$).

Discussion

One of the hallmarks of a scientific approach to studying literature and literary reading is the understanding that the act of reading literature and knowledge about literature does not exist in a vacuum. Instead, it is always performed by an *individual reader*, reading a *unique text*, in a *specific environment*, for a *certain purpose*. In large part, how the reader will approach the text depends on their previous experience. The same holds when it comes to discussing what literature is and what makes a text literary: It is the product of a mind that learned something from its contact with, i.e., literature.

Learning, in this sense, can be understood as a change in an organism's (be it an animal or a human) perception and/or conception of its environment (see Divjak, Milin, & Medimorec, 2020). From this perspective, the goal of learning is adaptive – it reduces uncertainty about the environment (see our opening quote, and also see Rescorla, 1988). The results of the present study shed further light on two crucial properties of learning that shape performance on genre decision tasks: Learning is an inevitable consequence of *experience* that is always *contextualized* or situated in an environment (i.e., a learning background or ambient).

Over three simulations in which such learning took place, we manipulated the amount and diversity of learning and fed it into arguably the simplest available algorithm for modelling cue competition in learning (Rescorla & Wagner, 1972; Widrow & Hoff, 1960; for a more recent discussion see Milin et al., 2020). The results of these simulations indicated that the accrual of experience is necessary for establishing the baseline predictability of the environment (also, validity or accuracy). Furthermore, the characteristics of the context in which such experience takes place modulate the competition between cues by, de facto, making them more or less salient, i.e., more or less predictable for some specific outcome (also, reliability or precision).

The simulations led to predictions that perhaps appeared counter-intuitive: Literature students, with their more diverse experience, will exhibit less distinctive (or more overlapping) poeticity judgements. At the same time, STEM students, with only a basic experience of poetry, will more readily provide distinct poeticity judgements.

The simulation-based predictions were, subsequently, tested in a simple genre (poetry) decision experiment. The predicted effect was further 'enhanced' by blocking the content of a text and not allowing purposeful reading. Arguably, genre decision happens immediately before *any* reading begins and, thus, represents a crucial factor in optimizing skilled reading (cf., Hanauer, 1996, 2001; Zwaan, 1991, 1994, etc.). From this perspective, our manipulation served the purpose of isolating this mechanism. This, however, is not to say that actual reading and the effect of content will not have the strongest effect of all. Its magnitude and/or speed are fascinating issues on their own but remain beyond the scope of this study.

The results of the experimental part of the study confirmed the aforementioned predictions: Literature students show more insignificant differences across experimental conditions than STEM students do. Strikingly so, this was reflected even at the level of their individual (i.e., by-participant) variation. The present findings are, thus, relevant for poeticity

studies and for studies of literature, more generally. They reveal the important role that learning plays in these uniquely human tasks and experiences. Furthermore, it seems that learning serves as a missing functional link between the studies of literature, on the one hand, and studies of expert behaviour, on the other hand.

In addition to the first point, genre decision turned out to be not at all that simple and straightforward for literary experts-in-making (i.e., Literary students). The diverse contexts and nuanced cues they have experienced make clear-cut genre classification challenging, if not altogether wrong or meaningless.

In addition to the second point, it is worth emphasising, once again, that the observed pattern of effects is predicted by mechanisms that are, from the perspective of high-level cognition, *domain general*. It follows that the implications reach far beyond the domain of text genre classification. In particular, the findings highlight the strong effects that experience can have on cognitive performance, even in apparently over-learned tasks. This parallels the findings of Bilalić, McLeod, and Gobet (2008) on expert chess-players who may, depending on the particularities of the task at hand, perform equally poorly as average players, and ever more so the more skilled they are (also see Bilalić, Langner, Erb, & Grodd, 2010). From a learning stance, these findings could be reinterpreted as evidence of how the amount of experience, jointly with the diversity of contexts, shapes consequent performance.

In a sense, the present findings highlight a truism that has been curiously neglected in studies of human cognition: Learning is always contextually situated, and when it comes to performance, what is learned can 'hurt' as well as 'help'; this has been demonstrated across a range of tasks from language (Ramscar, Sun, Hendrix, & Baayen, 2017) to spatial orientation (Woollett & Maguire, 2009, 2011). The context and its diversity enter into cue competition, simplifying (clearing) or complicating (blurring) the predictability of a situation and, as a consequence, performance in it (viz. 'failing expertise' in studies of Abernethy & Russell, 1987; Bilalić et al., 2014, and others). With that in mind, and until better evidence accumulates, we must refrain from the practice of assuming that 'worse' on a test means 'cognitively worse,' because once experience is controlled for, the opposite appears to hold.

What does this mean for genres? For one, it makes genres learned. Literary texts are, as much as any other object in the natural world, categorized to make dealing with them easier. This categorization is, almost counter-intuitively, more important to the layperson than to the expert, as experts may feel rather confined than assisted by categorization and are more likely to encounter examples that find themselves *in between genres*, or texts that happen to have a characteristic that would otherwise serve as an excellent cue to some other genre. In terms of literary theory, this probably places our perspective close to the one presented by Adam and Heidmann (2004), where genres are seen as ever-changing, with every new text altering the pool of texts to which it 'belongs.' The author, the publisher, and the reader all have knowledge of 'genericity' in literature and rely on that knowledge to write, distribute, and read literary texts, further altering the experience of what a certain genre represents. Whether any salient clusters of texts can be generated is perhaps a question better posed to corpus investigations of literary texts (but bear in mind that a human operator would have to select which texts are literary, and then use certain indices of group membership). The central point of this paper is that genres are learned from relatively limited experience in order to make dealing with texts easier, and a person's understanding of genre changes with new experience, to better suit their needs.

Conclusion

Most, if not all, human cognitive skills are learned in cue-rich contexts. Since experience continuously shapes this learning, it follows that experience and context can both serve to change behaviour on performance tests. This is true even of seemingly simple tasks, such as the genre decision task, the overlearned status of which becomes apparent in the speed and confidence exhibited by the raters. Yet, their responses can vary markedly depending on the quantity and quality of their experience.

The apparent contradiction of Literature students struggling more to make a resolute genre decision received a simple and elegant explanation in the learning framework. This links further to questions of general expertise and the conditions under which it can 'fail.' In fact, learning sheds light on what that 'failure' actually means, and why it becomes inevitable that any expert reader would, in time and with expanded experience, become disillusioned and claim that genre categories are void. No wonder, then, that literary scholars react against prescriptive notions of firmly demarcated literary genres (De Geest & Van Gorp, 1999; Derrida & Ronell, 1980; Miller & Kelly, 2016; Todorov, 2000).

Author contributions:
PM conceived and designed the experiment, analysed the data and interpreted the results, ran the computational simulations and wrote and critically revised the manuscript, with particular attention for framing the role of context in learning. FN collected the data but also contributed to the design of the experiment and to the interpretation of the results, and wrote those parts of the Introduction and Discussion relating to poeticity, genre categorization and genre decision making. MR contributed to framing the role of cue competition in learning, and to the discussion of the links between expertise and poetry and literary genre categorization; he also wrote parts of earlier versions of the Introduction and Discussion.

Acknowledgements:
We are greatly indebted to Dagmar Divjak (University of Birmingham, UK) and Kris Van Heuckelom (K.U. Leuven, Belgium) for their extensive feedback on earlier versions of this manuscript. We also wish to thank the two journal reviewers for their detailed comments. Last but not least, our gratitude goes to the journal's editors for their encouragement: to David Hanauer (Editor-In-Chief) and to Paul Sopčák (Assistant Editor).

Open Practices Statement:
The experimental data and the accompanying code for running the GAMMs presented in this study can be found at the following links

Code: https://github.com/ooominds/Experiencing_text_genre
Data: https://doi.org/10.25500/edata.bham.00000521

References

Abernethy, B., & Russell, D. G. (1987). The relationship between expertise and visual search strategy in a racquet sport. *Human Movement Science, 6*(4), 283-319. doi:10.1016/0167-9457(87)90001-7

Adam, J.-M., & Heidmann, U. (2004). Des genres à la généricité. L'exemple des contes (Perrault et les Grimm). *Langages*(1), 62-72.

Ambridge, B., & Lieven, E. V. M. (2011). *Child language acquisition : contrasting theoretical approaches*.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review, 118*(3), 438-481.

Bilalić, M., Grottenthaler, T., Nägele, T., & Lindig, T. (2014). The faces in radiological images: fusiform face area supports radiological expertise. *Cerebral Cortex, 26*(3), 1004-1014.

Bilalić, M., Langner, R., Erb, M., & Grodd, W. (2010). Mechanisms and neural basis of object and pattern recognition: a study with chess experts. *Journal of Experimental Psychology: General, 139*(4), 728-742.

Bilalić, M., McLeod, P., & Gobet, F. (2008). Why good thoughts block better ones: The mechanism of the pernicious Einstellung (set) effect. *Cognition, 108*(3), 652-661.

Blough, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes, 1*(1), 3-21.

Bortolussi, M., & Dixon, P. (2003). *Psychonarratology: Foundations for the empirical study of literary response*. Cambridge: Cambridge University Press.

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological bulletin, 114*(1), 80-99.

Bouton, M. E. (1997). Signals for whether versus when an event will occur. In M. E. Bouton & M. S. Fanselow (Eds.), *Learning, motivation, and cognition: The functional behaviorism of Robert C. Bolles* (pp. 385-409). Washington, DC: American Psychological Association.

Bradshaw, T., & Nichols, B. (2004). *Reading At Risk: A Survey of Literary Reading in America (Research Division Report #46)*. Retrieved from Washington, DC:

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology, 4*(1), 55-81.

Chen, Z., Haykin, S., Eggermont, J. J., & Becker, S. (2008). *Correlative learning: a basis for brain and adaptive systems*. New Jersey: John Wiley & Sons.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology, 18*(2), 185-196.

De Geest, D., & Van Gorp, H. (1999). Literary genres from a systemic-functionalist perspective. *European Journal of English Studies, 3*(1), 33-50.

Derrida, J., & Ronell, A. (1980). The law of genre. *Critical inquiry, 7*(1), 55-81.

Divjak, D., Milin, P., & Medimorec, S. (2020). Construal in language: A visual-world approach to the effects of linguistic alternations on event perception and conception. *Cognitive Linguistics, 31*(1), 37-72.

Enquist, M., Lind, J., & Ghirlanda, S. (2016). The power of associative learning and the ontogeny of optimal behaviour. *Royal Society open science, 3*(11), 160734.

Fox, J. (2002). *An R and S-Plus companion to applied regression*. Thousand Oaks, CA: Sage.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological review, 107*(2), 289-344.

Hanauer, D. (1996). Integration of phonetic and graphic features in poetic text categorization judgements. *Poetics, 23*(5), 363-380.

Hanauer, D. (1998a). The genre-specific hypothesis of reading: Reading poetry and encyclopedic items. *Poetics, 26*(2), 63-80.

Hanauer, D. (1998b). Reading poetry: An empirical investigation of formalist, stylistic, and conventionalist claims. *Poetics Today*, 565-580.

Hanauer, D. (2001). What we know about reading poetry. Theoretical positions and empirical research. In D. Schram & G. Steen (Eds.), *The psychology and sociology of literature* (pp. 107-128). Amsterdam: John Benjamins

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*(2), 96-101.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological review, 95*(4), 528-551.

Hoffstaedter, P. (1987). Poetic text processing and its empirical investigation. *Poetics, 16*(1), 75-91.

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory*. New York: Appleton-Century-Crofts.

James, W. (1909). *The meaning of truth*. New York: Longmans, Green and Company.

Kakade, S., & Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psychological review, 109*(3), 533-544.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering, 82*(1), 35-45.

Kimmel, H. D. (1959). Amount of conditioning and intensity of conditioned stimulus. *Journal of Experimental Psychology, 58*(4), 283-288.

Konorski, J. (1967). *Integrative activity of the brain; an interdisciplinary approach*. Chicago: University of Chicago Press.

Kruschke, J. K., & Hullinger, R. A. (2010). Evolution of attention in learning. In N. A. Schmajuk (Ed.), *Computational models of conditioning* (pp. 10-52). Cambridge: Cambridge University Press.

Landeira-Fernandez, J. (1996). Context and Pavlovian conditioning. *Brazilian Journal of Medical and Biological Research, 29*(2), 149-173.

Linderholm, T. (2006). Reading with purpose. *Journal of College Reading and Learning, 36*(2), 70-80.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior research methods, 44*(2), 314-324.

Miall, D. S. (2002). Literary discourse. In A. Graesser, M. Gernsbacher, & S. Goldman (Eds.), *Handbook of discourse processes* (pp. 321-355). New Jersey: Lawrence Erlbaum

Miall, D. S., & Kuiken, D. (1998). The form of reading: Empirical studies of literariness. *Poetics, 25*(6), 327-341.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PloS one, 12*(2), e0171935.

Milin, P., Madabushi, H. T., Croucher, M., & Divjak, D. (2020). Keeping it simple: Implementation and performance of the proto-principle of adaptation and learning in the language sciences. *arXiv, arXiv:2003.03813*.

Miller, C. R., & Kelly, A. R. (2016). Discourse genres In A. Rocci & L. d. Saussure (Eds.), *Verbal communication* (pp. 269-286). Berlin: Walter de Gruyter

Myers, C. E., & Gluck, M. A. (1994). Context, conditioning, and hippocampal rerepresentation in animal learning. *Behavioral neuroscience, 108*(5), 835-847.

Nadel, L., & Willner, J. (1980). Context and conditioning: A place for space. *Physiological Psychology, 8*(2), 218-228.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology, 53*(3), 139-154.

Peskin, J. (1998). Constructing meaning when reading poetry: An expert-novice study. *Cognition and Instruction, 16*(3), 235-263.

Peskin, J. (2007). The genre of poetry: Secondary school students' conventional expectations and interpretive operations. *English in Education, 41*(3), 20-36.

Peskin, J. (2010). The development of poetic literacy during the school years. *Discourse Processes, 47*(2), 77-103.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ramscar, M., Sun, C. C., Hendrix, P., & Baayen, H. (2017). The mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the “cost” of learning, not cognitive decline. *Psychological Science, 28*(8), 1171-1179.

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science, 31*(6), 927-960.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science, 34*(6), 909-957.

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific studies of reading, 10*(3), 241-255.

Reichle, E. D., & Reingold, E. M. (2013). Neurophysiological constraints on the eye-mind link. *Frontiers in Human Neuroscience, 7*, 361.

Rescorla, R. A. (1970). Reduction in the effectiveness of reinforcement after prior excitatory conditioning. *Learning and Motivation, 1*(4), 372-381.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American psychologist, 43*(3), 151-160.

Rescorla, R. A., Durlach, P. J., & Grau, J. W. (1985). Contextual learning in Pavlovian conditioning. In P. D. B. A. Tomie (Ed.), *Context and learning* (pp. 23-56). New Jersey: Lawrence Erlbaum.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Rosas, J. M., Aguilera, J. E. C., Álvarez, M. M. R., & Abad, M. J. F. (2006). Revision of retrieval theory of forgetting: What does make information context-specific? *international Journal of Psychology and Psychological therapy, 6*(2), 147-166.

Rosas, J. M., Todd, T. P., & Bouton, M. E. (2013). Context change and associative learning. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(3), 237-244.

Spence, K. W. (1956). *Behavior theory and conditioning*. New Haven: Yale University Press.

Spreat, S., & Spreat, S. R. (1982). Learning principles. *The Veterinary clinics of North America. Small animal practice, 12*(4), 593-606.

Sutton, R. S. (1992). *Gain adaptation beats least squares.* Paper presented at the Proceedings of the 7th Yale workshop on adaptive and learning systems.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497-537). Cambridge, MA: The MIT Press.

Todorov, T. (2000). The origin of genres. In D. Duff (Ed.), *Modern genre theory*. Harlow: Longman.

Tolman, E. C. (1949). There is more than one kind of learning. *Psychological review, 56*(3), 144-155.

Tomie, A. (1981). Effect of unpredictable food on the subsequent acquisition of autoshaping: Analysis of the context blocking hypothesis. In C. M. Locurto, H. S. Terrace, & J. Gibbon (Eds.), *Autoshaping and conditioning theory* (pp. 181-215). New York: Academic Press.

Trimmer, P. C., McNamara, J. M., Houston, A. I., & Marshall, J. A. (2012). Does natural selection favour the Rescorla–Wagner rule? *Journal of theoretical biology, 302*, 39-52.

Urcelay, G. P., & Miller, R. R. (2014). The functions of contexts in associative learning. *Behavioural processes, 104*, 2-12.

Van den Broek, P., Lorch, R. F., Linderholm, T., & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition, 29*(8), 1081-1087.

Van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2015). itsadug: Interpreting time series and autocorrelated data using GAMMs (Version 1.0.1): R statistical group.

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits.* Paper presented at the WESCON Convention Record Part IV.

Wood, S. (2007). The mgcv package: R statistical group.

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association, 111*(516), 1548-1563.

Woollett, K., & Maguire, E. A. (2009). Navigational expertise may compromise anterograde associative memory. *Neuropsychologia, 47*(4), 1088-1095.

Woollett, K., & Maguire, E. A. (2011). Acquiring "the Knowledge" of London's layout drives structural brain changes. *Current biology, 21*(24), 2109-2114.

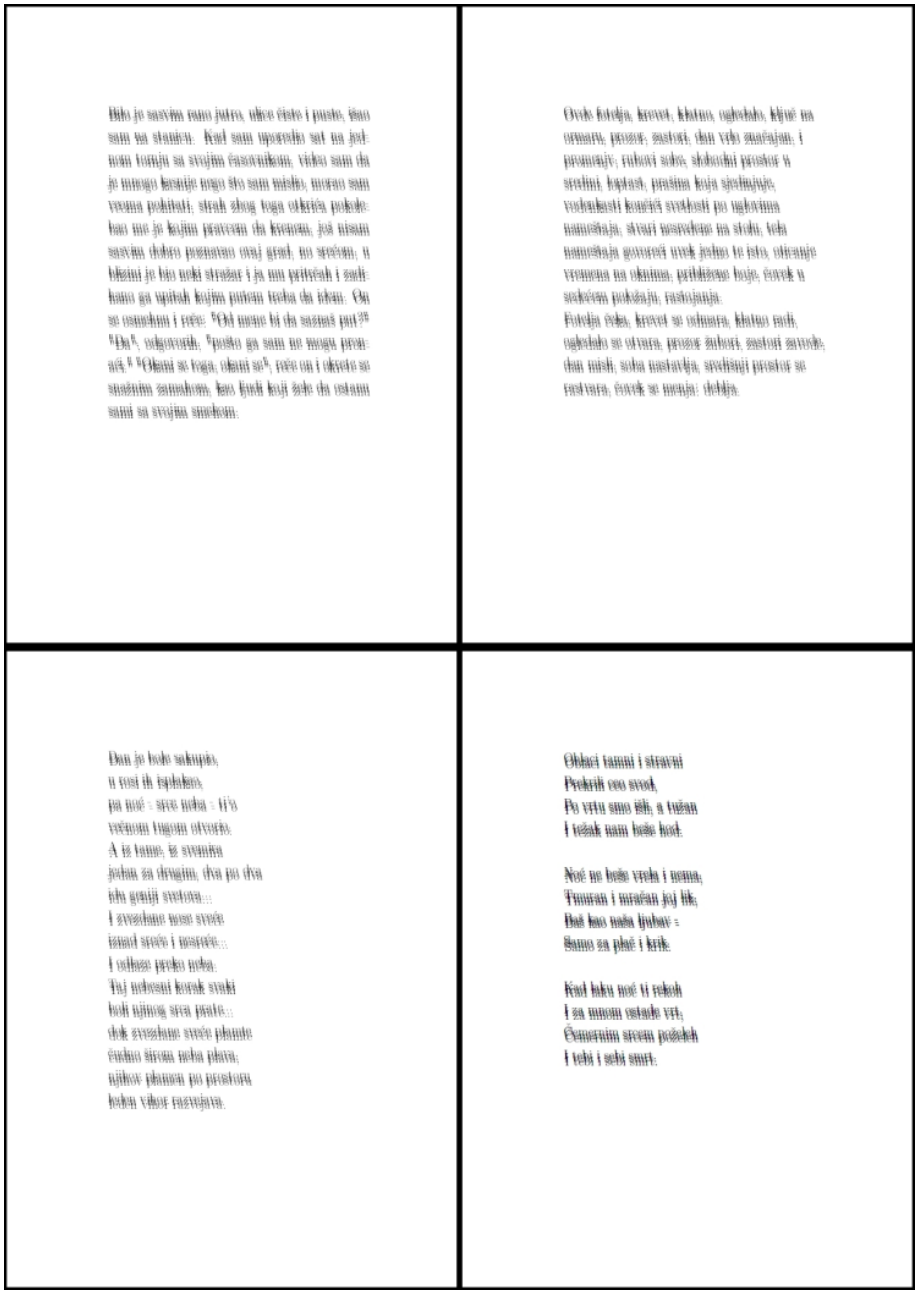
Zwaan, R. A. (1991). Some parameters of literary and news comprehension: Effects of discourse-type perspective on reading rate and surface structure representation. *Poetics, 20*(2), 139-156.

Zwaan, R. A. (1993). *Aspects of literary comprehension: A cognitive approach* (Vol. 29). Amsterdam: John Benjamins.

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 920-933.

Appendices

Appendix A:
Sample of blurred texts used as stimuli in the behavioural experiment, one in each of the four graphical layouts (not to scale).



Appendix B:
Summary tables of basic descriptive statistical indicators.

Table B.1. *Medians (Mdn) and Median Absolute Deviations (MAD) for the ratings on "To what extent is the text a poem?" question.*

Experience	Text source	Text form							
		Stanzas		Verses		Left		Double	
		Mdn	MAD	Mdn	MAD	Mdn	MAD	Mdn	MAD
STEM	**Anthology**	3	0.00	2	1.48	-1	1.48	-2	1.48
	R. Book	3	0.00	2	1.48	-1	1.48	-2	1.48
	Article	2	1.48	1	1.48	-2	1.48	-2	1.48
	Text	3	0.00	2	1.48	-2	1.48	-2	1.48
Literature	**Anthology**	3	0.00	3	0.00	0	2.97	-1	2.97
	R. Book	3	0.00	2	1.48	-2	0.00	-3	1.48
	Article	3	0.00	2	1.48	-2	0.00	-3	1.48
	Text	3	0.00	2	1.48	-2	1.48	-2	1.48

Table B.2. *Percentage of YES-answers to "Is the text a poem or not?" question.*

Experience	Text source	Text form			
		Stanzas	Verses	Left	Double
STEM	**Anthology**	99.38	86.25	33.13	20.63
	R. Book	98.13	82.50	23.75	17.50
	Article	88.75	69.38	13.13	5.00
	Text	96.88	81.88	18.75	11.88
Literature	**Anthology**	98.75	94.38	46.88	38.75
	R. Book	97.50	86.25	20.63	8.75
	Article	92.50	83.13	25.00	11.25
	Text	98.75	93.75	26.88	14.38

Table B.3. *Medians (Mdn) and Median Absolute Deviations (MAD) for the ratings on "How confident are you in your previous ratings?" question.*

Experience	Text source	Text form							
		Stanzas		Verses		Left		Double	
		Mdn	MAD	Mdn	MAD	Mdn	MAD	Mdn	MAD
STEM	**Anthology**	5	0.00	4	1.48	4	1.48	4	1.48
	R. Book	5	0.00	4	1.48	4	1.48	4	1.48
	Article	5	0.00	4	1.48	4	1.48	5	0.00
	Text	5	0.00	4	1.48	4	1.48	4	1.48
Literature	**Anthology**	5	0.00	5	0.00	4	1.48	5	0.00
	R. Book	5	0.00	5	0.00	4	1.48	5	0.00
	Article	5	0.00	5	0.00	4	1.48	5	0.00
	Text	5	0.00	5	0.00	4	1.48	4	1.48

Appendix C:
Summary tables of statistical analyses: Generalized Additive Mixed Modelling (GAMM). The presented models all have the better fit statistics in comparison with simpler models.

Table C.1. *Generalized additive mixed model fitted to the ratings on the question "To what extent is the text a poem?", using the ordered categories link-function (ocat) with 7 levels. Reporting parametric coefficients (Part A) and effective degrees of freedom (edf), reference degrees of freedom (Ref.df), F and p values for the random effects (Part B).* $Deviance\ explained = 48.1\%$.

Parametric coefficients	**Estimate**	**St. Error**	**z-value**	**Pr (>|z|)**
[Reference level; i.e., Intercept]				
Stanzas Anthology STEM	5.869	0.324	18.137	< 0.0001
Verses Anthology STEM	-2.287	0.380	-6.021	< 0.0001
Left-aligned Anthology STEM	-5.293	0.380	-13.929	< 0.0001
Justified Anthology STEM	-5.946	0.381	-15.609	< 0.0001
Stanzas R.Book STEM	-0.207	0.252	-0.822	0.411
Stanzas Article STEM	-1.261	0.252	-4.998	< 0.0001
Stanzas Text STEM	-0.433	0.252	-1.722	0.085
Stanzas Anthology Literature	1.580	0.421	3.757	0.0002
Verses Anthology Literature	0.507	0.461	1.100	0.271
Left-aligned Anthology Literature	-0.737	0.460	-1.603	0.109
Justified Anthology Literature	-0.962	0.462	-2.082	0.037
Stanzas R.Book Literature	-1.268	0.362	-3.508	0.0005
Stanzas Article Literature	-0.484	0.362	-1.338	0.181
Stanzas Text Literature	-0.823	0.361	-2.283	0.022
Smooth terms	**Edf**	**Ref.df**	**Chi.sq**	**p-value**
Participant & Form	195.11	248	2426.00	< 0.0001
Participant & Source	168.18	248	1480.00	< 0.0001
Item	65.18	76	632.00	< 0.0001

Number of observations: 5120; Participants: 64; Items: 80

Table C.2. *Generalized additive mixed model fitted to the answers on "Is the text a poem or not?" question using binomial link-function, reporting parametric coefficients (Part A) and effective degrees of freedom (edf), reference degrees of freedom (Ref.df), F and p values for the random effects (Part B).* $Deviance\ explained = 66.3\%$.

Parametric coefficients	**Estimate**	**St. Error**	**z-value**	**Pr (>|z|)**
[Reference level; i.e., Intercept]				
Stanzas Anthology STEM	4.660	0.404	11.531	< 0.0001
Verses Anthology STEM	-1.840	0.379	-4.860	< 0.0001
Left-aligned Anthology STEM	-5.651	0.379	-14.932	< 0.0001
Justified Anthology STEM	-6.605	0.386	-17.127	< 0.0001
Stanzas R.Book STEM	-0.444	0.384	-1.155	0.248
Stanzas Article STEM	-1.641	0.388	-4.235	< 0.0001
Stanzas Text STEM	-0.782	0.384	-2.038	0.042
Stanzas Anthology Literature	1.065	0.405	2.630	0.009
Stanzas R.Book Literature	-1.326	0.546	-2.431	0.015
Stanzas Article Literature	-0.111	0.547	-0.204	0.839
Stanzas Text Literature	-0.471	0.544	-0.865	0.387
Smooth terms	**Edf**	**Ref.df**	**Chi.sq**	**p-value**
Participant & Form	110.44	251	451.10	< 0.0001
Participant & Source	166.55	248	752.60	< 0.0001
Item	55.29	76	385.10	< 0.0001

Number of observations: 5120; Participants: 64; Items: 80

Table C.3. *Generalized additive mixed model fitted to the ratings on "How confident are you in your previous ratings?" question using the ordered categories link-function (ocat) with 5 levels. Reporting parametric coefficients (Part A) and effective degrees of freedom (edf), reference degrees of freedom (Ref.df), F and p values for the random effects (Part B).* $Deviance\ explained = 27.2\%$.

Parametric coefficients	**Estimate**	**St. Error**	**z-value**	**Pr (>|z|)**
[Reference level; i.e., Intercept]				
Stanzas Anthology STEM	5.877	0.293	20.039	< 0.0001
Verses Anthology STEM	-1.566	0.235	-6.670	< 0.0001
Left-aligned Anthology STEM	-1.873	0.235	-7.980	< 0.0001
Justified Anthology STEM	-1.452	0.236	-6.154	< 0.0001
Stanzas R.Book STEM	0.007	0.143	0.047	0.963
Stanzas Article STEM	-0.285	0.143	-1.998	0.046
Stanzas Text STEM	-0.150	0.143	-1.052	0.293
Stanzas Anthology Literature	1.221	0.391	3.121	0.002
Verses Anthology Literature	0.167	0.311	0.537	0.592
Left-aligned Anthology Literature	-0.548	0.309	-1.775	0.076
Justified Anthology Literature	-0.475	0.311	-1.528	0.127
Smooth terms	**Edf**	**Ref.df**	**Chi.sq**	**p-value**
Participant	51.76	62	4143.10	< 0.0001
Participant & Form	122.13	248	586.80	0.009
Participant & Source	121.47	251	500.90	0.003
Item	46.49	76	149.00	< 0.0001

Number of observations: 5120; Participants: 64; Items: 80