Clustering Hand-Drawn Sketches
via Analogical Generalization

Maria D. Chang, Kenneth D. Forbus

Qualitative Reasoning Group, Northwestern University

2133 Sheridan Road, Evanston, IL 60208

maria.chang@u.northwestern.edu, forbus@northwestern.edu

Abstract

One of the major challenges to building intelligent educational software is determining what kinds of feedback to give learners. Useful feedback makes use of models of domain-specific knowledge, especially models that are commonly held by potential students. To empirically determine what these models are, student data can be clustered to reveal common misconceptions or common problem-solving strategies. This paper describes how analogical retrieval and generalization can be used to cluster automatically analyzed hand-drawn sketches incorporating both spatial and conceptual information. We use this approach to cluster a corpus of hand-drawn student sketches to discover common answers. Common answer clusters can be used for the design of targeted feedback and for assessment.

Introduction

Sketching and drawing are valuable tools for reasoning about space. Creating visualizations and diagrams lightens working memory load and makes spatial inference easier (Larkin & Simon, 1987). Sketching is especially useful for learning and instruction in spatially rich subjects, like science, technology, engineering and mathematics (i.e. STEM fields). Indeed, spatial skills have been identified as a predictor of STEM success (Wai, Lubinski & Benbow, 2009). For science education, sketching can be used to increase engagement, improve learning and encourage encoding across different representations (Ainsworth, Prain & Tytler, 2011).

Advances in intelligent tutoring systems have opened the possibility of creating educational software than can support sketching and take advantage of the many benefits it has to offer (Valentine et al. 2012). However, building

intelligent sketching software is challenging. The noisy nature of sketches makes them difficult to interpret. Consequently, assessing the quality of a student's sketch requires a considerable amount of spatial and conceptual reasoning. With the exception of advanced design sketches, most sketches are rough approximations of spatial information. They are rarely drawn to scale and often require multimodal cues (e.g. gestures, speech) to facilitate understanding. For example, a sketch of a map might contain various shapes that represent different landmarks. The shapes may look nothing like the actual landmarks, but may be denoted as landmarks by labels. No one has trouble understanding that a blob can represent something that looks physically different; such visual information is processed with a grain of salt. Somehow, people are able to focus on the spatially and conceptually important information in the sketch and, for the most part, ignore irrelevant information.

Qualitative representations are a good match for sketched data because they carve continuous visual information (e.g. 2D location) into discrete categories and relationships (e.g. round, right of, etc). These representations enable software systems to reason about sketches using the same structured representations that are hypothesized to be salient for humans.

Since sketches are not always understood in isolation, a model of visual comparison that also incorporates conceptual information is also important for sketch understanding. Analogical comparison using structure-mapping (Gentner 1983) allows structured descriptions to be compared to each other to evaluate how similar the two descriptions are. The structure-mapping model of analogy has built in constraints and biases that are supported by psychological research. Computational models of structure mapping have been used in cognitive simulation (Gentner & Forbus, 2011; Lovett & Forbus, 2011), physics problems

solving (Klenk & Forbus, 2009; Lockwood & Forbus 2009), and question answering (Murdock, 2011). Structure-mapping enabled these systems to make more human-like comparisons. In educational software, such as Sketch Worksheets (Yin et al. 2010) structure-mapping generates comparisons that can be used to assess a student's sketch by comparing it to a pre-defined solution.

A major challenge in any intelligent tutoring system is determining how to coach students. When designing feedback, instructors must hypothesize what will be hard for students. Such hypotheses are not always data driven and can be inaccurate (Nathan, Koedinger & Alibali, 2001). Consequently, most successful intelligent tutoring systems incorporate detailed cognitive models of the task being taught. Building cognitive models requires research on novice misconceptions and strategies (Anderson et al. 1995). Some systems also model the strategies of human tutors, such as intervention techniques and tutoring dialogue (VanLehn et al. 2007). However, creating cognitive models for both correct knowledge and common misconceptions for an entire domain is difficult. On the other hand, specific exercises can have easily defined misconceptions that can be identified without a full analysis of the domain. By analyzing the work of multiple students on an example, common models (some of which may be misconceptions) can be mined from the data. Although there has been work devoted to assessing student knowledge through sketches (Jee et al. 2009; Kindfield, 1992) and mining information about students from learning data (e.g. from hand-coded sketches, Worsley & Bliknstein, 2011) we are unaware of any efforts to combine automatic sketch understanding and educational data mining. This paper describes an approach for using analogical reasoning over hand-drawn sketches to detect common student answers.

Our hypothesis is that analogical generalization can be used to generate meaningful clusters of hand-drawn sketches. We compare analogical generalization to a k-means clustering algorithm and evaluate its performance on a set of labeled (i.e. clustered by hand) student sketches. The resulting clusters from the experiments can be inspected to identify the key characteristics of each cluster. These characteristics can be used to identify student misconceptions and to design targeted feedback for students.

Background

Structure-mapping is the comparison mechanism of our clustering approach. Here we summarize the computational models for analogical matching, retrieval and generalization that we use. We then describe Sketch Worksheets, which is our sketch-based educational software system used to collect and encode hand-drawn sketches.

Structure Mapping

The structure-mapping engine (SME) (Faulkenhainer, Forbus & Gentner, 1989) is a computational model of analogy that compares two structured descriptions, a *base* and a *target*, and computes one or more analogical mappings between them. Each mapping contains a set of *correspondences*, which indicate which items in the base correspond to which items in the target, a *structural evaluation score*, which is a measure of match quality, and a set of *candidate inferences*, which are statements that are true in the base and hypothesized to be true in the target. Several constraints are imposed on the mapping process to prevent all potential mappings from being computed and to account for certain psychological phenomena. Importantly, SME has a bias for mappings with greater *systematicity*, which means that it prefers mappings with systems of shared relations.

To create sketch clusters, we use SAGE (*sequential analogical generalization engine*), an extension of SEQL (Kuenhe et al 2000) which computes probabilities for expressions during generalization and retrieves structured descriptions using analogical retrieval. Generalizations are created by incrementally introducing exemplars into a *generalization context*. Each generalization context consists of a case library that includes both exemplars and generalizations. For each new exemplar, the most similar exemplar or generalization in the generalization context is retrieved by an analogical reminding via MAC/FAC (Forbus, Gentner & Law, 1995). MAC/FAC computes *content vectors* that measure the relative frequency of occurrence of relations and attributes in structured representations. It finds the maximum dot product of the vector for the probe with the vectors of everything in the generalization context. This step may retrieve up to three items and is analogous to a bag of words approach to similarity, albeit with predicates. These items are then compared to the probe using SME and the item with the highest structural evaluation score is returned.

The reminding returned by MAC/FAC (either an exemplar or an existing generalization) is merged with the new exemplar if their similarity is above a pre-defined *assimilation threshold*. If the best match is a generalization, the new exemplar is added to it. If the best match is another exemplar, then the two are combined into a new generalization. If there is no best match above the assimilation threshold, then the new exemplar is added directly to the case library for that generalization context. It will remain an exemplar in the generalization context until it is joined with a new exemplar or until there are no more exemplars to add.

The resulting generalizations contain generalized facts and entities. Each fact in a generalization is assigned a probability, which is based on its frequency of occurrence in the exemplars included in the generalization. For example, a fact that is true in only half of the exemplars would be assigned a probability of 0.5. Thus entities become more abstract, in that facts about them "fade" as their probability becomes lower.
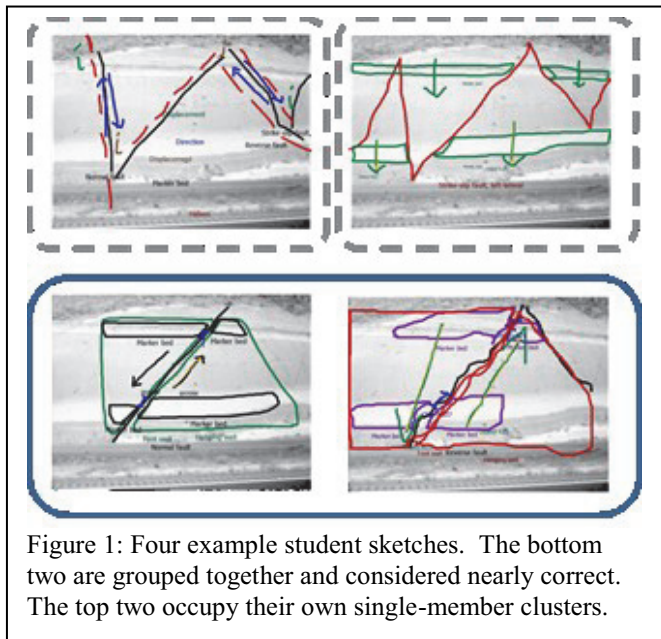

Figure 1: Four example student sketches. The bottom two are grouped together and considered nearly correct. The top two occupy their own single-member clusters.

Sketch Worksheets

Sketch Worksheets are built within CogSketch (Forbus et al 2011), our open-domain sketch understanding system. Each sketch worksheet includes a problem statement, a pre-defined solution sketch and a workspace where the student sketches his or her candidate answer. As part of the authoring process, the worksheet author describes the problem and sketches an ideal solution. CogSketch analyzes the solution sketch by computing qualitative spatial and conceptual relations between items in the sketch. The worksheet author can then peruse these representations and identify which facts are important for capturing the correctness of the sketch. The author can also identify which drawn elements have quantitative location criteria by defining *quantitative ink constraints*, which define a tolerance region for a particular drawn element. If the student's drawn element falls outside of the tolerance region, it is considered incorrect. If it falls within the tolerance region, it is considered correct.

The difference between these two criteria types can be illustrated by two different worksheet exercises. Consider a worksheet that asks a student to draw the solar system. The exact location of the sun does not matter, as long as it is contained by the orbit rings of other planets. In other words, its location is constrained by its location relative to other drawn entities. Alternatively, consider a worksheet that asks a student to identify the temporal lobe on a diagram of the human brain. The absolute location of the drawing denoting the temporal lobe would be important. For an element whose location is constrained relative to an absolute frame of reference (e.g. a background image), quantitative ink constraints are necessary.

Sketch Worksheets have been used in experiments on spatial reasoning as well as classroom activities in geoscience and elementary school biology. The sketches used in the experiments described in this paper were collected using sketch worksheets.

Clustering via Analogical Generalization

Clustering is achieved by performing analogical generalization over student sketches. The clustering algorithm adds the sketches in random order, using the SAGE algorithm described above. A single generalization context is used, i.e. it operates unsupervised, because the goal is to see what clusters emerge.

Encoding

A major challenge to clustering sketches is choosing how to encode the information depicted in each sketch. Each sketch contains a wealth of spatial information, not all of it relevant for any particular situation.[1] In order to highlight visually and conceptually salient attributes and relationships, we harness information explicitly entered by the student and the worksheet author. More specifically, we filter the underlying representations in each sketch based on the following principles: conceptual information is critical, quantitative ink constraints must constrain analogical mappings, and worksheet authoring should guide spatial and conceptual elaboration.

Conceptual Information
Every sketch worksheet comes equipped with a subset of concepts from an OpenCyc-derived knowledge base. This subset contains the concepts that may be used in the worksheet and are selected by the worksheet author to limit the conceptual scope of the exercise. These concepts are applied by students to elements of their drawing via CogSketch's conceptual labeling interface. This is useful for education because the mapping between shapes and entities is often one to many. While visual relationships are computed automatically by CogSketch, conceptual relationships are entered by sketching arrows or

---

[1] For example, a student might draw a planet above the sun versus below the sun, a visually salient difference that doesn't matter in most orbital diagrams.

annotations and labeling them appropriately, via the same interface. Thus the conceptual labels constitute the student's expression of their model of what is depicted. Consequently, conceptual information is always encoded for generalization.

Quantitative Ink Constraints Limit Matches

Another type of information that is entered explicitly by the worksheet author are quantitative ink constraints. Recall that quantitative ink constraints define a tolerance region relative to an absolute frame of reference (e.g. a background image). Quantitative ink constraints are defined for entities whose absolute position matters.

When encoding information about entities for which there are quantitative ink constraints, the encoding algorithm computes their position with respect to the tolerance regions, to determine if the entity's location meets the constraint or not. If it does not, we further encode how the constraint was violated (e.g. too wide, too narrow, etc.) and include that information in the encoding.

Furthermore, each entity that is evaluated with respect to a quantitative ink constraint is associated with that constraint as a location-specific landmark. This association limits the possible analogical mappings by ensuring that entities associated with one landmark cannot map to entities that are associated with a different landmark. This also ensures that entities cannot be generalized across different location-specific landmarks. This approach for using quantitative constraints to limit the analogical mappings has been shown to lead to sketch comparisons that provide more accurate feedback to students (Chang & Forbus, 2012).

Spatial and Conceptual Elaboration

Worksheet authors can also specify a subset of the visual relationships computed by CogSketch as important. For example, the core of the Earth must be inside its mantle. Some conceptual information can also be marked as important, e.g. a sketch of a fault must include marker beds. All facts marked as important by the worksheet author, whether spatial or conceptual, are always included in the encoding for generalization.

Evaluation

To evaluate our clustering algorithm we used a set of fault identification worksheets (e.g. Figure 1) submitted by students taking an undergraduate geoscience course at Northwestern University. There were 28 sketches in total, spanning three different fault identification exercises. A gold standard was created by hand-clustering the sketches for each exercise separately. We then ran our generalization algorithm on the unlabeled data for each exercise, to evaluate how well the clusters it produced match the gold standard. Because clusters may differ depending on the order in which sketches are selected, we repeated the clustering over 10 iterations. We collected three measures from the resulting clusters: purity, precision and recall.

K-Means Clustering

To explore the impact of relational structure on generalization behavior, we also compared our approach to a non-structural way of ascertaining similarity. Specifically, we used the MAC/FAC content vectors (described above) as a cruder, non-relational form of similarity. While content vectors are still sensitive to the presence of relationships, since those predicates are included in them, it only contains relative frequency information. In other words, "man bites dog" is the same as "dog bites man." We used k-means clustering on the same data, where each mean was the content vector of a sketch and the distance measure between means was the inverse dot product of the content vectors being compared. The more overlap between the content vectors, the greater the similarity and the smaller the distance. For each k-means clustering process we supplied k by counting the number of labeled clusters. In this sense, the k-means clustering approach had a slight advantage over analogical generalization. The k-means clustering algorithm was also repeated 10 times, since the initial k means can impact the makeup of clusters.

Results

Table 1 shows the average purity, precision and recall for each approach across the 3 worksheet groups, averaged over 10 iterations of each approach. Analogical generalization outperformed k-means without analogy for clustering in all measures. Since purity is often high when there are many clusters, it is important to consider the precision and recall measures as well.

We used independent samples t-tests to test for significant differences between purity, precision and recall for each sketch group separately (for a total of 9 comparisons). Each measure was significantly higher for analogical generalization than for k-means clustering ($p < 0.005$, Bonferroni corrected for 9 comparisons).

Figure 2 shows two sketches that were frequently generalized together. This cluster indicates a common sketching behavior exhibited by students. The high probability facts in the generalization indicate the defining criteria for the cluster. Most of the high probability facts in this generalization are concept membership attributes. Other facts refer to the direction of the sketched diagonal arrows in the sketch. These facts were already considered in the feedback design of this worksheet. However, the three high probability facts shown in Figure 2 indicate the potential for more targeted feedback. These facts indicate that three of the four marker beds failed quantitative ink constraints in specific ways. The bold horizontal arrows imposed on the figure point to two marker beds that map to each other in an analogical mapping. Both of these marker beds fall short of the left bounds of their quantitative ink constraints (see first fact in Figure 2). Similarly, two other marker beds (unmarked) fall short of the right bounds of the quantitative ink constraints. Without knowing that multiple students would exhibit this common behavior, a worksheet author would have no reason to include targeted feedback about it. However, given that multiple students commit this error, targeted feedback about the horizontal extent of marker beds would have been helpful, e.g. "Marker bed regions are not just near the fault; they can extend to the edges of the image."


```
(lessThanQuantInkLeftBound
   (GenEntFn 5 1 Faults-1) Object-438)
(lessThanQuantInkRightBound
   (GenEntFn 3 1 Faults-1) Object-440)
(lessThanQuantInkRightBound
   (GenEntFn 4 1 Faults-1) Object-436)
```
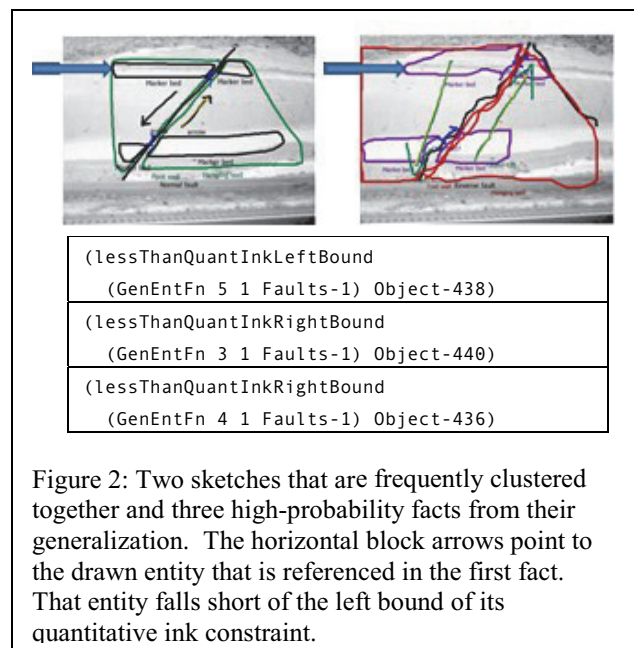
Figure 2: Two sketches that are frequently clustered together and three high-probability facts from their generalization. The horizontal block arrows point to the drawn entity that is referenced in the first fact. That entity falls short of the left bound of its quantitative ink constraint.

	SAGE	k-means
Sketch Group #1		
Number of Clusters	6.7	6
Purity**	0.90	0.72
Precision**	0.86	0.56
Recall*	0.85	0.56
Sketch Group #2		
Number of Clusters	5.5	4
Purity**	0.94	0.74
Precision**	0.98	0.61
Recall**	0.82	0.59
Sketch Group #3		
Number of Clusters	6.1	4
Purity**	0.96	0.83
Precision**	0.99	0.67
Recall*	0.80	0.68

Table 1: Clustering measures for analogical generalization (SAGE) and k-means clustering (without analogy). All measures are averaged over 10 random restart iterations of the clustering procedure. Asterisks indicate the probability associated with independent samples t-tests between SAGE and k-means measures: ** $p < 0.001$, * $p < 0.005$.

Related Work

Many researchers have explored misconceptions in domains like algebra, geometry (Anderson et al. 1995) and physics (VanLehn et al. 2007). Each of these research programs answers important questions about the structure of knowledge during leaning. These answers have shaped the coaching strategies of various tutoring systems.

Many sketch understanding systems exist but most stick to a single domain because they use sketch recognition (Lee et al. 2007; de Silva et al. 2007; Valentine et al. 2012). No other sketch understanding systems use structure-mapping as a model for comparison. Despite this, it may still be possible to apply similar clustering techniques to those systems.

Discussion and Future Work

This paper describes a method for clustering sketches to detect common answer patterns. We used models of human analogical processing to cluster hand-drawn sketches completed by undergraduate geoscience students. The analogical clustering approach significantly outperformed a k-means clustering algorithm.

This technique can be used to mine common answer patterns from sketches so that they can be used for assessment or for designing targeted feedback. Instructors may use this technique to discover the distribution of answer patterns in their classrooms, some of which may be prevalent misconceptions. This approach enables common answer detection in a data-driven (but tightly scoped)

manner, without requiring a cognitive analysis of the entire domain or even the entire task.

One of the limitations to this approach is the understandability of the facts used to describe generalizations. As discussed above, high-probability facts can be used to understand the defining criteria of a cluster. For an instructor to easily interpret these facts would require familiarity with the knowledge representations used there. However, it can be argued that the instructor may not need those explicit facts. Instead, they can simply view a prototypical member of the cluster and decide on the defining criteria for themselves. With this technique, rather than looking at all the sketches submitted by students, an instructor can inspect only as many sketches as there are clusters.

In the future we plan to continue refining encoding procedures of sketches. The procedures used in this experiment are domain general, but there are likely cases where tighter filters on conceptual information will be needed. We also have not yet integrated shape and edge level representations into this encoding procedure (Lovett et al. 2012), as these are only now starting to be integrated into our sketch worksheets. We also plan to add clustering to the grading utilities built into sketch worksheets.

Acknowledgements

References

Ainsworth, S., Prain, V. and Tytler, R., 2011. Science education. Drawing to learn in science. *Science*, 333(6046), 1096-1097.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R., 1995. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2), 167-207.

Chang, M.D., Forbus, K.D., 2012. Using quantitative information to improve analogical matching between sketches. Innovative Applications of Artificial Intelligence (IAAI). Toronto, Canada.

de Silva, R., Bischel, T. D., Lee, W., Peterson, E. J., Calfee, R. C. and Stahovich, T., 2007. Kirchhoff's pen: A pen-based circuit analysis tutor. In Proceedings of the 4th Eurographics workshop on Sketch-based interfaces and modeling.

Falkenhainer, B., Forbus, K., & Gentner, D., 1989. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1), 1-63.

Forbus, K., Gentner, D., and Law, K., 1995. MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19, 141-205.

Forbus, K. D., Usher, J., Lovett, A., Lockwood, K. and Wetzel, J., 2011. Cogsketch: Sketch understanding for cognitive science research and for education. *Topics in Cognitive Science*, 3, 648-666.

Gentner, D. & Forbus, K., 2011. Computational models of analogy. *Wiley Interdisciplinary Reviews: Cognitive Science*. 2(3), 266-276.

Gentner, D., 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155-170.

Jee, B., Gentner, D., Forbus, K. D., Sageman, B. and Uttal, D. H., 2009. Drawing on experience: Use of sketching to evaluate knowledge of spatial scientific concepts. Proceedings of the 31st Annual Conference of the Cognitive Science Society.

Kindfield, A. C. H., 1992. Expert diagrammatic reasoning in biology. In Proceedings of the AAAI Spring Symposium. Palo Alto, CA.

Klenk, M. and Forbus, K., 2009. Analogical model formulation for transfer learning in AP Physics. *Artificial Intelligence* 173(18), 1615-1638.

Kuehne, S., Forbus, K., Gentner, D. and Quinn, B., 2000. SEQL: Category learning as progressive abstraction using structure mapping. Proceedings of CogSci 2000.

Larkin, J. H. and Simon, H. A., 1987. Why a diagram is (sometimes) worth 10000 words. *Cognitive Science*, 11(1), 65-99.

Lee, W., De Silva, R., Peterson, E. J., Calfee, R. C. and Stahovich, T. F., 2007. Newton's pen – a pen-based tutoring system for statics. In Proceedings of the Eurographics Workshop on Sketch-Based Interfaces and Modeling. Riverside, CA.

Lockwood, K. & Forbus, K., 2009. Multimodal knowledge capture from text and diagrams. Proceedings of KCAP-2009.

Lovett, A., Kandaswamy, S., McLure, M., & Forbus, K., 2012. Evaluating qualitative models of shape representation. Proceedings of the 26th International Workshop on Qualitative Reasoning. Los Angeles, CA.

Murdock, J., 2011. Structure mapping for jeopardy! clues. Case-Based Reasoning Research and Development, 6-10.

Nathan, M., Koedinger, K. R., & Alibali, M., 2001. Expert blind spot: When content knowledge eclipses pedagogical content knowledge. In Proceedings of the Third International Conference on Cognitive Science, 644-648.

Valentine, S., Vides, F., Lucchese, G., Turner, D., Rim, H., Li, W., Linsey, J., Hammond, T., 2012. Mechanix: A Sketch-Based Tutoring System for Statics Courses. Innovative Applications of Artificial Intelligence (IAAI). Toronto, Canada.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P., 2007. When are tutorial dialogues more effective than reading?. *Cognitive Science*, 31(1), 3-62.

Yin, P., Forbus, K. D., Usher, J., Sageman, B. and Jee, B., 2010. Sketch worksheets: A sketch-based educational software system. Innovative Applications of Artificial Intelligence (IAAI). Portland, OR.

Wai, J., Lubinski, D., & Benbow, C. P., 2009. Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, 101(4), 817.

Worsley, M. and Blikstein P., 2011. What's an Expert? Using learning analytics to identify emergent markers of expertise through automated speech, sentiment and sketch analysis. In Proceedings for the 4th Annual Conference on Educational Data Mining.