Using Student Writing and Lexical Analysis to Reveal Student Thinking about the Role of Stop Codons in the Central Dogma

Luanna B. Prevost,[†]* Michelle K. Smith,[‡] and Jennifer K. Knight[§]

[†]Department of Integrative Biology, University of South Florida, Tampa, FL 33620; [‡]School of Biology and Ecology and Maine Center for Research in STEM Education, University of Maine–Orono, Orono, ME 04469; [§]Department of Molecular, Cellular, and Developmental Biology, University of Colorado Boulder, Boulder, CO 80309

ABSTRACT

Previous work has shown that students have persistent difficulties in understanding how central dogma processes can be affected by a stop codon mutation. To explore these difficulties, we modified two multiple-choice questions from the Genetics Concept Assessment into three open-ended questions that asked students to write about how a stop codon mutation potentially impacts replication, transcription, and translation. We then used computer-assisted lexical analysis combined with human scoring to categorize student responses. The lexical analysis models showed high agreement with human scoring, demonstrating that this approach can be successfully used to analyze large numbers of student written responses. The results of this analysis show that students' ideas about one process in the central dogma can affect their thinking about subsequent and previous processes, leading to mixed models of conceptual understanding.

INTRODUCTION

The topic of information flow has been cited by many groups as a fundamental biological concept (e.g., the *Vision and Change in Undergraduate Education* document [American Association for the Advancement of Science, 2011] and the *Next Generation Science Standards* [NGSS, 2013]), yet students at both the secondary and postsecondary levels have difficulty understanding how genetic information is stored and exchanged (e.g., Lewis *et al.*, 2000; Mills Shaw *et al.*, 2008). Students also display conceptual difficulties about the related subtopic of the central dogma, the process by which genetic information, DNA, is encoded into proteins via an RNA intermediate. For example, students often believe that the amino acids used in translation are manufactured by the translation process itself (Fisher, 1985). In addition, students often fail to accurately explain the relationship between DNA and protein or between genes and proteins (Wood-Robinson, 2000; Marbach-Ad, 2001; Mills Shaw *et al.*, 2008). Students' misunderstandings about these relationships are further revealed in their ideas about the impact of mutations on transcription and translation (Smith *et al.*, 2008; Smith and Knight, 2012).

A strong conceptualization of the central dogma is important for understanding basic cell functions and the origin and expression of genetic disorders, yet several barriers potentially confound students. For example, biologists typically summarize the central dogma with diagrams such as: DNA → RNA → protein. This visual representation may cause difficulties for students when they do not understand the processes represented by each arrow (Wright *et al.*, 2014). Even students who do understand the meaning of each arrow may not understand the multiple steps within each process of

Jennifer Momsen, *Monitoring Editor*

Submitted December 30, 2015; Revised May 23, 2016; Accepted May 25, 2016

CBE Life Sci Educ December 1, 2016 15:ar65

DOI:10.1187/cbe.15-12-0267

*Address correspondence to: Luanna B. Prevost (prevost@usf.edu).

© 2016 L. B. Prevost *et al.* CBE—Life Sciences Education © 2016 The American Society for Cell Biology. This article is distributed by The American Society for Cell Biology under license from the author(s). It is available to the public under an Attribution–Noncommercial–Share Alike 3.0 Unported Creative Commons License (http://creativecommons.org/licenses/by-nc-sa/3.0).

"ASCB®" and "The American Society for Cell Biology®" are registered trademarks of The American Society for Cell Biology.

Supplemental Material can be found at:
http://www.lifescied.org/content/suppl/2016/11/29/15.4.ar65.DC1

Downloaded from http://www.lifescied.org/ by guest on February 15, 2018

replication, transcription, and translation, or how the processes relate to one another. Ultimately, diagnosing student difficulties on concepts such as these is the first step in understanding how to help students repair their incorrect ideas and develop a deeper understanding of the processes involved.

To this end, many concept inventories have been developed to address a variety of known difficult topics (e.g., Garvin-Doxas, 2008; Kalas *et al.,* 2013; Price *et al.*, 2014). Two concept inventories in particular, the Genetics Concept Assessment (GCA; Smith *et al.*, 2008) and the Genetics Literacy Assessment Instrument (Bowling *et al.*, 2008), specifically address genetic information flow and the potential effects of mutations. Studies on student conceptual change and on learning progressions have also demonstrated that students seldom hold only novice ideas or only expert scientific ideas (Chi *et al.*, 1981; Vosniadou and Brewer, 1992); instead, both novice and scientific ideas can coexist in students' minds. Multiple-choice assessments, in which students are restricted to selecting only one option, reveal whether students hold one scientific or one novice idea but do not capture multiple ideas that students may hold. To understand the details of student thinking on these topics, instructors need more sensitive tools that allow measurement of the sequence of student thinking and that indicate whether students hold both correct and incorrect ideas simultaneously.

Constructed-response assessments, in which students answer questions in their own words, can be used to reveal more nuanced levels of student understanding, thus providing instructors with deeper insight into student thinking (Birenbaum and Tatsuoka, 1987; Martinez, 1999; Kuechler and Simkin, 2010). For example, constructed-response assessments can be used to determine whether students' mental models use purely scientific concepts, only novice ideas, or a combination of both scientific and novice ideas (Opfer *et al.*, 2012). Additionally, writing gives students practice in scientific discourse, which not only helps them to communicate their pre-existing ideas but to construct new ideas about scientific phenomena (Lemke, 1990; Keys, 1999). Despite the advantages of written assessments, the considerable time and effort required to read, assess, and provide feedback on student responses limits their use by instructors. Automating the process of analyzing student responses can greatly reduce this barrier (Nehm and Haertig, 2012), and recent innovations in lexical analysis and machine learning have enabled successful analysis of student writing on a variety of biological concepts (Ha *et al.*, 2011; Haudek *et al.*, 2012; Weston *et al.*, 2015).

In this paper, we build on the previously reported result from the GCA—that students fail to understand the effects of a stop codon mutation on transcription and translation (Smith *et al.*, 2008; Smith and Knight, 2012)—by developing open-ended questions modified from GCA multiple-choice items. We then apply lexical analysis to examine whether student written responses can provide additional insight into their thinking. By following sequential student responses to each of three questions about the effect of a stop codon mutation on replication, transcription, and translation, we explore whether students have incorrect ideas about each of these processes, or whether their ideas about one process affect their descriptions of subsequent or previous processes.

METHODS

In this section, we describe the questions we developed and a novel approach to understanding student ideas that combines lexical and statistical analyses to examine student responses.

Questions and Administration

The GCA consists of 25 multiple-choice questions designed to assess nine learning goals in undergraduate genetics courses. This study examines one of the learning goals of the GCA: compare different types of mutations and describe how each can affect genes and the corresponding mRNAs and proteins. Two GCA questions that address this learning goal ask students to identify which of several mutations would result in a shortened mRNA, and which would result in a shortened polypeptide (Figure 1A). For the first GCA question, none of the changes will result in a shortened RNA, but 47% of majors and nonmajors students ($n = 352$ students) from multiple classes selected the stop codon mutation as their answer (Smith and Knight, 2012). For the second GCA questions, 35% of students did not answer that a frameshift mutation can produce a stop codon, which would lead to a shorter protein. For the work described in this paper, we modified these questions to create three open-ended prompts that ask students to explain how a mutation that results in a premature stop codon would affect replication, transcription, and translation (Figure 1B).

We administered the three open-ended questions to undergraduate biology majors enrolled in introductory cell and molecular biology–focused courses at two large public research universities. Demographic data for students in these courses are included in Table 1. The questions were administered as part of a regular online homework assignment after the instructors had completed teaching a unit on central dogma. Students were encouraged to give their best effort and received one point for completion of the entire assignment. We collected responses from 1043 students who answered all three open-response questions.

Lexical Analysis

We used IBM (2011b) SPSS Text Analysis for Surveys and IBM (2011a) SPSS Modeler to analyze the responses to each question. Before this analysis, student responses were preprocessed. Preprocessing included removal of HTML characters and spell-checking. Student answers were uploaded to the lexical analysis software. Words and short phrases were extracted using preprogrammed libraries that came with the software package and custom libraries that were built to capture words commonly used in biology courses (for more details, see Haudek *et al.*, 2012; Kaplan *et al.*, 2014). For creation of new custom libraries, words that were not available in the preprogrammed libraries were added to a new library. Commonly misspelled words were also added to the library so that, if these words were left uncorrected in these or future responses, the misspelled words could be recognized by the software. For the software to correctly match the misspelling to the correctly spelled word, these words were grouped as synonyms. For example, the misspelled word "transcripted" was added as a synonym of "transcribed." Extracted words and short phrases were then grouped into categories. A category contains all words including synonyms and phrases in students writing that express a homogeneous idea. Categories could also contain

Downloaded from http://www.lifescied.org/ by guest on February 15, 2018

A GCA multiple-choice questions

Use the following mRNA codon key as needed to answer the next two questions:

GCC Alanine	GAA Glutamic acid
AAU Asparagine	GAG Glutamic acid
CCU Proline	AGG Arginine
GGA Glycine	CCC Proline
UGG Tryptophan	CAU Histidine
UGA "Stop" (no amino acid)	

The following DNA sequence (coding strand) occurs near the middle of the coding region of a gene.

DNA

50 55 60 65

5'– A A T G A A T G G G A G C C T G A A G G A G –3'

The corresponding mRNA sequence is shown below. Note that the coding strand of DNA has the same sequence as the mRNA, except that there are U's in the mRNA where there are T's in the DNA. The first triplet of nucleotides AAU (underlined) is in frame for coding, and encodes Asparagine as the codon table above indicates.

mRNA

50 55 60 65

5'– A A U G A A U G G G A G C C U G A A G G A G –3'

1. Which of the following DNA mutations is almost certain to result in a shorter than normal mRNA?

a) A→G at position 50
b) G→A at position 53
c) C→A at position 58
d) None of the above

2. For the same DNA sequence, which of the following DNA mutations is almost certain to result in a shorter than normal protein?

a) T→C at position 50
b) A→ at position 61
c) Insertion of a G after the G at position 54
d) None of the above

B Open-response questions

The following DNA sequence occurs near the middle of the coding region of a gene
DNA 5' A A T G A A T G G* G A G C C T G A A G G A 3'

There is a G to A base change at the positon marked with an asterisk. Consequently, a codon normally encoding an amino acid becomes a stop codon.

a) How will this alteration influence DNA replication?
b) How will this alteration influence transcription?
c) How will this alteration influence translation?

FIGURE 1. (A) Original question from the GCA (Smith et al., 2008). Correct answers in bold. (B) Open-response stop codon questions used in this study.

Boolean characters (e.g., AND, NOT, OR) to create detailed phrases that express specific ideas.

Throughout this paper, categories will be represented in italics to help identify them. For example, both of the responses in Table 2 were assigned to the *replication* category. The *replication* category was represented by the phrase "replication process" in the first example response and "DNA replication" in the second response. Both responses were also assigned to the *replication unaffected* category. The first response contained the phrase "will not have any effect" and the second the phrase "will not

Downloaded from http://www.lifescied.org/ by guest on February 15, 2018

TABLE 1. Demographic data from the two introductory cell and molecular–focused courses

Demographic data		Institution 1 (%)	Institution 2 (%)
Gender			
	Female	51	56
	Male	49	44
Ethnicity			
	Native American/Pacific Islander	2	<1
	Asian	9	7
	Black	3	8
	Hispanic	8	4
	White	74	80
	Other	4	2
Year in school			
	First year	68	13
	Second year	18	61
	Junior	6	17
	Senior	5	8
	Other	3	1
Major			
	Biology	53	52
	Other arts and sciences	35	29
	Engineering	7	11
	Other	5	8

influence," both of which convey the homogeneous idea of not having an effect on replication. The second response in Table 2 also was assigned to the categories *stop codon* and *translation*. Each student response can be assigned to zero, one, or multiple categories based on the words and phrases it contains. In Table 2, the first response is assigned to two categories, and the second is assigned to four categories.

Response Coding

The authors coded the same subset of 211 responses to each of the three questions using a holistic rubric in which each response was characterized as a 1, 2, or 3. Examples of student responses in each bin are given in Table 3. We have presented student responses exactly as they were written but have made a few edits to correct misspellings or grammar in order to make responses easier to read.

For the replication and transcription responses, the following rubric was used to categorize each student response into bins:

TABLE 2. Responses to the replication question and their assigned lexical categories

Response[a]	Category: Replication	Replication unaffected	Stop codon	Translation
It will not have any effect on the replication process	√	√		
It will not influence DNA replication as a new stop codon will only affect translation.	√	√	√	√

[a]All terms in categories are underlined and highlighted in colored text that corresponds to the category to which they belong.

1. Correct: no effect on replication or transcription
2. Incomplete/irrelevant: contains some but not all correct information or irrelevant response (e.g., discussing translation)
3. Incorrect: replication or transcription stops

For the translation responses, the rubric was

1. Correct: translation stops
2. Incomplete: contains some but not all correct information
3. Incorrect: no effect on translation

The raters achieved an interrater reliability (IRR) of 0.8 or higher in scoring responses to each of the three questions for the initial 211 responses. We calculated the intraclass correlation (Cronbach's alpha), which can be used to compare agreement among more than two raters. Values of 0.7 and higher are considered acceptable levels of IRR (Shrout and Fleiss, 1979; Cronbach, 1984; Crocker and Algina, 1986). Because acceptable IRR levels were achieved, each rater was assigned one question; one rater (J.K.K.) coded the remaining replication responses, the second rater (M.K.S.) coded the remaining transcription responses, and the third rater (L.B.P.) coded the remaining translation responses.

Model Building using Multinomial Logistic Regression

We used all the categories developed during the lexical analysis and expert coding to build predictive models to automatically score new responses. Because answers from students at each institution displayed a comparable distribution of categories with similar frequencies, we pooled the data from the two institutions. We randomly selected 70% of the responses as a training set ($n = 730$) and built one model for each question using the training set. The remaining 30% of the responses were reserved to test each model ($n = 313$).

We used multinomial logistic regression to build each model. In our analyses, the dependent variable was the expert coding for each question, which includes values of 1, 2, or 3 (correct, incomplete, or incorrect, as described above). The multinomial logistic regression uses one value of the dependent variable as a reference. In this case, the reference is bin 1 (correct) from the human scoring. The independent variables in the model were the lexical analysis categories. Each lexical analysis category variable is binary and indicates the presence or absence of that category in a student's response. The multinomial logistic regression builds one model whose fit is determined using a Pearson chi-squared–based maximum-likelihood test (Menard, 2002). We used a stepwise regression with $F_{in} = 0.05$ and $F_{out} = 0.10$ criteria for inclusion and exclusion of each category in the model to determine the subset of the lexical analysis categories that were most predictive of human scoring.

The multinomial logistic regression model predicts the likelihood that a response will be classified as incomplete/irrelevant (bin 2) or incorrect (bin 3) compared with correct (bin 1, the reference value). For each question, one model is created. However, two sets of regression coefficients are generated for each comparison. One set of regression coefficients is used to predict the likelihood that a student's response would be classified as incomplete compared with correct, and the second set of regression coefficients is used to predict the likelihood that a student's response would be classified as incorrect compared with correct.

We evaluated the regression coefficients (β) to determine the contribution of each lexical category to the model logit. The

Downloaded from http://www.lifescied.org/ by guest on February 15, 2018

TABLE 3. Examples of student responses and rubric coding, with each response shown given by a different student.

	Bins		
Question	1: Correct	2: Incomplete/irrelevant	3: Incorrect
Replication	It will not have any effect on the replication process.	This would be an example of a nonsense mutation.	The DNA will stop replicating when it reaches the stop codon.
Transcription	It will not have any effect on transcription.	This will cause a mutation in the transcription process.	In the process of transcribing DNA into RNA, the newly added stop codon will inhibit the rest of the chain from being transcribed into RNA.
Translation	This will influence translation because the stop codon will cause the amino acid sequence to end before it should. This will create a different polypeptide or protein that will either not function or function differently than it should have.	The code will be translated with a different base and will be read differently. This will result in a different protein being built.	This will have no influence on translation.

logit is the natural log of the odds ratio for the model. We used the Wald statistic to evaluate the significance of the regression coefficients in the model, that is, whether the regression coefficient is different from zero or whether the odds ratio is different from correct (bin 1). In each model, the regression coefficient represents the degree to which that lexical analysis category would affect the model logit if its value increased from 0 to 1, holding other categories constant.

The exponent of the regression coefficient gives the odds ratio. Odds ratios vary from zero (0) to infinity (∞) and describe the odds of the comparison expert coding changes corresponding to the reference coding (correct or bin 1 in these analyses) when the category variable changes from 0 to 1 (Menard, 2002; Peng *et al.*, 2002). Thus, we present the likelihood that a category would increase or decrease the chance that a response is assigned to bin 2 compared with bin 1, or bin 3 compared with bin 1. We can illustrate how to interpret these odds using the incomplete/irrelevant response bin: the regression compares bin 2 with bin 1 (correct; the reference bin). An odds ratio of 1 indicates that it is equally likely that a score will be assigned to bin 2 or bin 1. An odds ratio greater than 1 indicates that it is more likely that the presence of a category would cause a response to be assigned to bin 2 (incomplete/irrelevant) than bin 1 (correct). Inversely, an odds ratio smaller than 1 indicates that it is less likely that the presence of a category would cause a response to be classified as incomplete/irrelevant rather than correct.

Finally, we compared the model's ability to correctly predict human scoring. The predictive models were tested using a test data set composed of 30% of the student responses. Models were deemed acceptable when they achieved an IRR between the computer prediction and human scoring of Cronbach's alpha 0.7 or greater.

RESULTS

Students provided written explanations for the effect of a base change in DNA on each of the processes of replication, transcription, and translation. The authors binned the student answers as 1, 2, or 3, as described in the *Response Coding* section of the *Methods*. Lexical analysis was used to identify categories of terms that students use in their writing about each process. The variables produced in the lexical and human scoring were then used to build a statistical model that can predict the human scoring of student responses. In addition, we identified any patterns in student answers that provided information about how students' understanding or misunderstanding of one process impacted their answers to subsequent questions.

Lexical Analysis

The lexical analysis identified sets of categories for each set of student answers: 34 categories for replication, 27 categories for transcription, and 24 categories for translation. All categories were used in our analyses; however, for ease of viewing, Figure 2 shows the most common categories for each question, grouped by the question in which they most frequently occur. For example, the categories *replication*, *replication unaffected*, and *replication stops* are unique to the replication question. Two categories, *DNA* and *stop codon*, are common to all three questions. The category *stop* is frequently used in responses to the translation and transcription questions.

Multinomial Logistic Regression Analysis

For each of the three questions (replication, transcription, and translation), we used the categories of word choices derived from the lexical analysis as predictors of expert human coding in a multinomial logistic regression analysis. Each predictive model achieved a human–computer IRR of 0.7 or greater for both the training and testing data sets (Cronbach's alpha; Table 4). Each model identified the subset of lexical analysis categories that best predict human expert coding.

For the replication question, the odds ratios for lexical analysis categories that best predict human expert coding are shown in Table 5. For this model, the incomplete/irrelevant (bin 2) and incorrect (bin 3) answers are compared with the correct (bin 1) category. For the incomplete/irrelevant bin, student responses are more likely to be classified as incomplete/irrelevant if they contain terms in the categories *transcription*, *amino acids*, *protein*, *replication stops*, and *stop codon* (odds ratios significantly greater than 1). Responses are less likely to be scored as incomplete/irrelevant if they contain terms in the categories *DNA*, *base*, *normal*, and *replication unaffected* (Table 5; odds ratios significantly less than 1). For the incorrect bin, student responses are more likely to be classified as incorrect if they contain terms in the categories *replication stops*, *stop codon*, *protein*, and *short* (odds ratios significantly greater than 1).

Downloaded from http://www.lifescied.org/ by guest on February 15, 2018

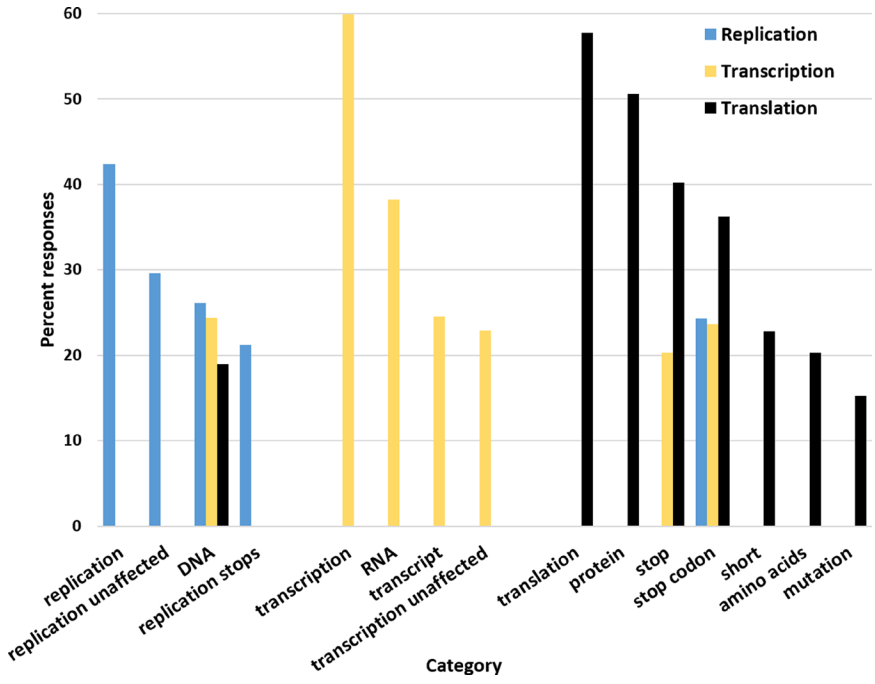

FIGURE 2. Distribution of responses by lexical analysis category for each question: replication (blue), transcription (yellow), and translation (black). The first four categories are the most commonly used terms for the replication question. The next four are most common to transcription, and the last seven are most common to translation. Some terms, such as "DNA," were used in answering all three questions. Only categories that garnered 15% or more of student responses are shown.

Responses are less likely to be scored as incorrect if they contain terms in the categories *replication unaffected*, *base*, *translation*, *protein*, *mutation*, and *DNA* (Table 5; odds ratios significantly less than 1). For brevity, we have included the odds ratios for the transcription and translation questions in Supplemental Tables S1 and S2.

Student Answers and Idea Maps

Once student responses were scored as 1, 2, or 3, the relationships between the terms students used were visualized (Figures 3, 4, and 5). These idea maps illustrate the proportion of responses (denoted by size or circle or node) within each bin that have been assigned to each lexical category. The connections between two categories represent the proportion of responses shared by these two categories. Solid lines connecting two categories show that more than 50% of the responses were shared by the two categories. Dashed lines connecting two categories indicate that 26–50% of the responses were shared by the two categories. Dotted lines connecting two categories show that 11–25% of the responses were shared by the two categories. Categories that share fewer than 11% of the responses share the categories are not connected. The idea maps for each of the three questions (replication, transcription, and translation) are described in more detail in the following sections.

Replication

Correct Responses (Bin 1). About half of the student responses to the replication question (56%) indicated that the base change had no effect on replication (Figure 3A). The frequency and association of the categories significant in predicting a correct response (bin1: "replication is unaffected") are displayed in Figure 3B. Only two categories, *replication unaffected* and *DNA*, have high frequencies for predicting correct responses, as represented by the large circles.

Incomplete or Irrelevant Responses (Bin 2). Twelve percent of the replication responses were categorized as incomplete or irrelevant. These responses varied widely and often did not give enough information to ascertain whether the student understood the effect of the mutation on replication. Some students did not write about the effect on replication. This shift away from describing replication is illustrated by the relative sizes of the categories shown in Figure 3C.

For example, many responses discussed transcription or translation but did not mention replication (67%). Smaller percentages of students talked generally about kinds of mutations (16%) or mentioned only that the DNA, RNA, or protein would be "different" (18%).

Incorrect Responses (Bin 3). About one-third (31%) of the replication responses were classified as incorrect. Responses were frequently categorized into *replication stops*, *DNA*, and *stop codon* (Figure 3D, larger circles), and these categories are associated with one another (dotted lines). Responses in this bin typically described the stop codon terminating the process of replication (Table 5).

TABLE 4. Evaluation of multinomial logistic regression models

Model	Pearson's goodness-of-fit chi-squared statistic[a]	Degrees of freedom	*p* Value[a]	Prediction accuracy[b] (Cronbach's alpha)
Replication	1480.6	802	<0.05	0.84
Transcription	1436.0	828	<0.05	0.74
Translation	1277.3	742	<0.05	0.70

[a]Goodness-of-fit statistics and *p* value evaluate whether there is a significant relationship ($p < 0.05$) between the dependent and independent variables.
[b]Accuracy values provided for testing data set $n = 313$.

Downloaded from http://www.lifescied.org/ by guest on February 15, 2018

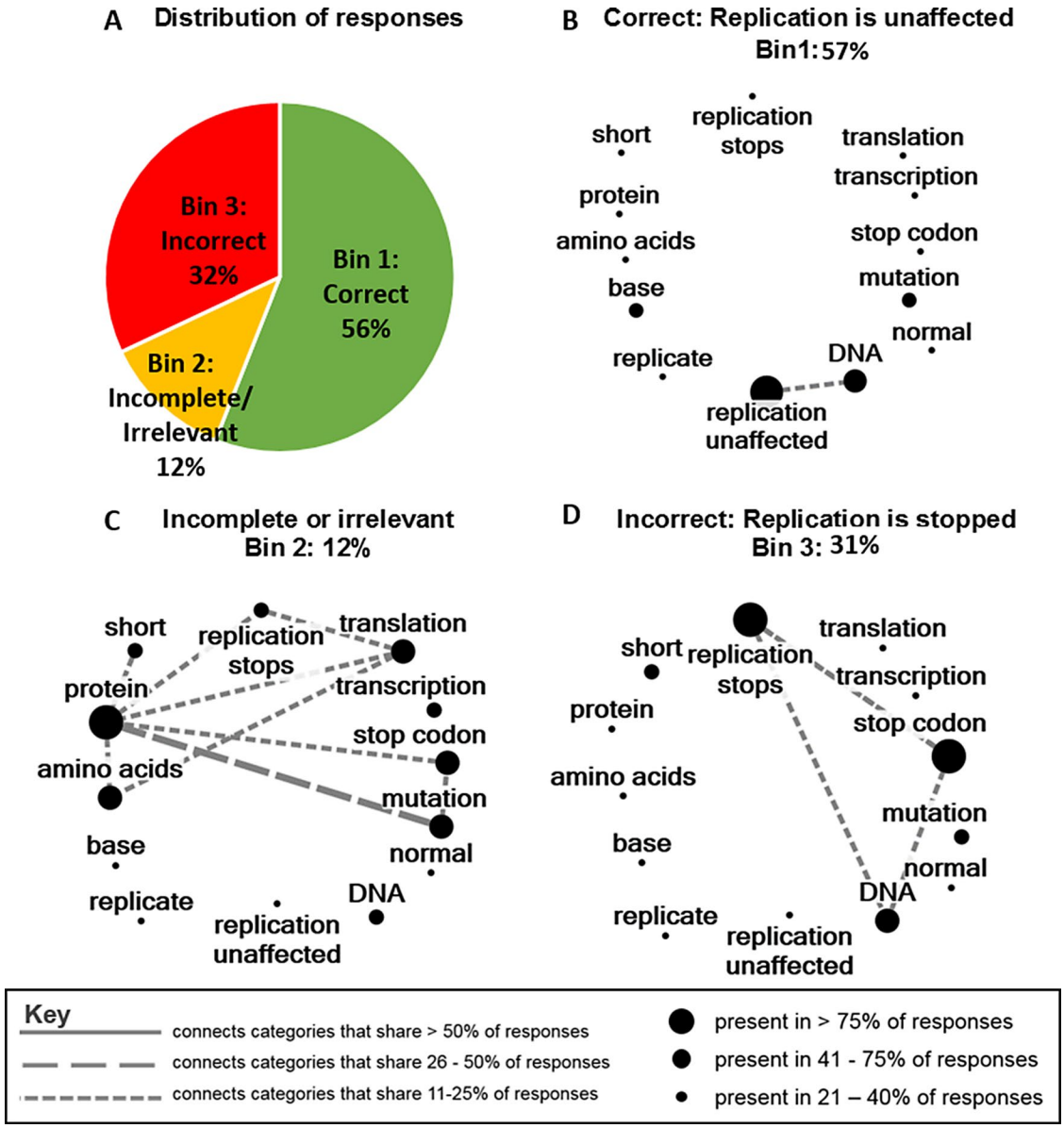

FIGURE 3. Student performance (A) and idea maps (B–D) for replication. Idea maps illustrate the frequency and cooccurrence of concepts in student writing.

Transcription

Correct Responses (Bin 1). Slightly more than half of the responses (55%) correctly identified that the base change would have no effect on transcription (Figure 4, A and B). The most frequent categories in correct responses are *mutation*, *translation*, *different*, *DNA*, and *no effect*. The associations between the categories *transcription* and *no effect*, and *transcription* and *DNA* (Figure 4B), reflect the two primary responses: 1) the alteration would not influence transcription, and 2) the alteration would cause a mutation in the DNA and/or the mRNA. Some students went on to describe that the complementary mRNA transcript would have a uracil base (U) instead of a cytosine (C) base.

Incomplete or Irrelevant Responses (Bin 2). Seven percent of responses to the transcription question fell into this category. Incomplete responses often indicated that something would go wrong during transcription without specifying the problems that may occur. Most responses (58%) were incomplete in some way, such as mentioning that there would be an impact on the transcription process without any detail about that change. Irrelevant responses mentioned other stages of the central dogma (37%) or discussed kinds of mutations (6%) rather than discussing transcription.

Incorrect Responses (Bin 3). Thirty-eight percent of the responses to the transcription question incorrectly stated that

Downloaded from http://www.lifescied.org/ by guest on February 15, 2018

TABLE 5. Odds ratios and confidence intervals for multinomial logistic regression for the open-response replication question[a]

	Incomplete/irrelevant (bin 2)[b]				Incorrect (bin 3)[b]			
Text analysis category	**β**	**OR**	**CI**	**Sig**	**β**	**OR**	**CI**	**Sig**
Transcription	2.02	7.57	(2.09–27.46)	↑	–0.10	0.90	(0.17–4.74)	NS
Amino acids	1.78	5.92	(2.00–17.49)	↑	–0.85	2.34	(0.75–7.24)	NS
Protein	1.40	4.07	(1.87–8.85)	↑	–1.32	0.27	(–0.10–0.74)	↓
Replication stops	1.17	3.22	(1.32–7.88)	↑	3.94	51.22	(23.28–112.7)	↑
Stop codon	0.83	2.30	(1.10–4.83)	↑	1.55	4.71	(2.39–9.28)	↑
DNA	–1.19	0.31	(0.16–0.58)	↓	–0.56	0.57	(0.31–0.78)	↓
Base	–1.96	0.14	(0.06–0.36)	↓	–2.43	0.09	(0.03–0.24)	↓
Normal	–2.32	0.10	(0.16–0.59)	↓	–1.24	0.29	(0.07–1.17)	NS
Replication unaffected	–4.30	0.1	(0.01–0.43)	↓	–2.98	0.05	(0.02–0.11)	↓
Mutation	0.41	1.50	(0.82–2.76)	NS	–0.74	0.48	(0.24–0.96)	↓
Translation	0.28	1.32	(0.59–2.94)	NS	–1.55	0.21	(0.08–0.57)	↓
Short	0.83	2.29	(0.73–7.20)	NS	2.79	16.34	(5.31–50.26)	↑
Replicate	–2.22	0.11	(0.1–1.04)	NS	–1.82	0.16	(0.03–1.00)	NS

[a]Reference value of the dependent variable score = correct.
[b]β, regression coefficient; OR, odds ratio; CI, 95% confidence interval; Sig, significance. ↑, Significant increase compared with reference value $p < 0.05$; ↓, significant decrease compared with reference value $p < 0.05$; NS, no significant change in odds ratio compared with reference value $p > 0.05$.

transcription stopped (Figure 4D). The categories *transcription*, *stop*, *transcript*, and *stop codon* were the most frequent. The strongest association occurred between *transcription* and *stop*. Incorrect responses included two primary, related incorrect ideas: 1) transcription would be stopped, and 2) the transcript or RNA product would be shorter because the alteration in the DNA strand produced a stop codon (Table 6).

Translation

Correct Responses (Bin 1). More than half of the students (57%) responded that translation stops and/or a shorter protein is produced (Figure 5, A and B). The most frequently used categories are *protein*, *stop*, *stop codon,* and *short* as shown by the larger circles in the idea map. These categories are also commonly used together in the same response. For example, more than 25% of the responses that were assigned to the category *stop* were also assigned to the category *protein* as demonstrated by the dashed line between these two categories in Figure 5B.

Incomplete or Irrelevant Responses (Bin 2). Twenty percent of responses to the translation question were classified as incomplete or irrelevant (Figure 5C). Most responses that were irrelevant (88%) described that translation will not occur properly or that the protein will function differently. Thus, some of the most frequent categories in this idea map were *protein*, *different*, *improper translation*, and *DNA*. Responses assigned to the category *DNA* referred to translation being different because of the change to the DNA sequence. Some responses categorized as incomplete were not detailed enough to ascertain whether the student knew that translation will be stopped due to the stop codon or that the protein formed would be shorter. Other responses discussed another stage of the central dogma rather than translation (9%). A few responses talked about mutations in general (3%).

Incorrect Responses (Bin 3). Twenty-three percent of students answered this question incorrectly (Figure 5D). The categories *DNA*, *mRNA*, *different*, *protein*, and *short* were most common among the answers. Of the student answers in this bin, about one-quarter (26%) stated that translation would not be altered by the creation of a stop codon. Another group of students indicated that translation would not occur at all (14%; Table 6). Other students (11%; Table 6) revealed a problem with replication or transcription that carried on to translation. For example, the base change would result in a shorter DNA strand during replication, but translation would proceed unaffected.

Tracing Student Thinking across Central Dogma Processes

By tracing the path students took when answering all three questions rather than focusing on their answers to individual questions, we were able to follow student thinking about how a mutation that produces a stop codon affects all three processes. Figure 6A shows the path taken by the 57% of students who answered the replication question correctly. Figure 6B shows the path taken by students who gave incomplete/irrelevant answers to the replication question. Finally, the paths taken by students who answered replication incorrectly are displayed in Figure 6C. We found that just more than one-quarter of students (26%) answered all three questions correctly, even though more than 50% of student responses were correct for any single question. This "all correct" path is illustrated in Figure 6A by the green circles and arrows. Fifty-seven percent of the students ($n = 586$) answered the replication question correctly. Of these students, 68% ($n = 401$) also answered the transcription question correctly. And of these 401 students who answered the replication and transcription questions correctly, 68% ($n = 272$) answered the translation question correctly. Conversely, ~5% of all students had completely incorrect explanations for all three questions. We obtained this percentage by following the red circles and arrows in Figure 6C. Thirty-one percent of students incorrectly answered the replication question, 61% of these students also answered the transcription question incorrectly, and 27% of this latter group answered the translation question incorrectly. Thus, most students displayed mixed mental models—combinations of correct and incorrect ideas across the three

Downloaded from http://www.lifescied.org/ by guest on February 15, 2018

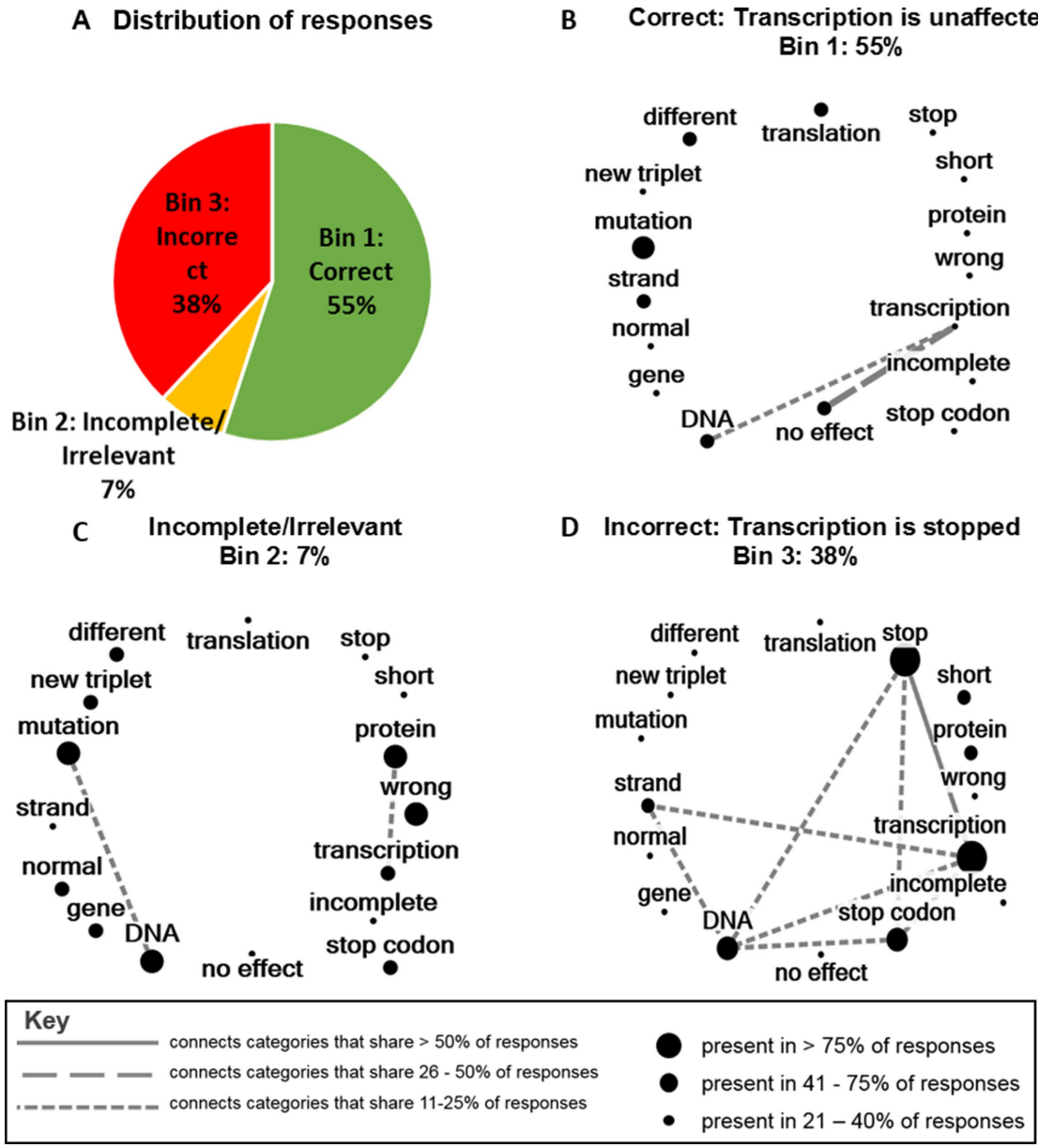

FIGURE 4. Student performance (A) and idea maps (B–D) for transcription. Idea maps illustrate the frequency and cooccurrence of concepts in student writing.

questions. These mixed models are represented by the additional pathways in Figure 6.

Mixed-Model Student Answers

There are several different ways in which students displayed mixed ideas, demonstrating correct understanding of one process but incorrect understanding of other processes. We have illustrated some of these mixed ideas below.

For example, students can answer translation correctly but answer both replication and transcription incorrectly. Strikingly, 51% of students who answered the replication and transcription questions incorrectly arrived at a correct explanation for translation (bin 3 → bin 3 → bin 1; Figure 6C, purple arrow). This group of students represents 10% of all students who responded to this suite of questions. Thus, 10% of all students correctly identified that translation would be stopped and/or the protein produced would be shorter despite incorrect ideas about the effect on both replication and transcription. These students thought that the base change would cause replication and transcription as well as translation to stop. For example, one student wrote: "This change to a stop codon will halt DNA replication and the genes that follow the stop codon will not be replicated. >> This stop codon will affect transcription because once again all of the DNA after the stop codon will not be transcribed into mRNA and therefore won't be expressed as genes. >> Translation will be affected because all of the mRNA codons after the stop codon will not be translated into its corresponding amino acid. Therefore important genetic information has been lost."

Downloaded from http://www.lifescied.org/ by guest on February 15, 2018

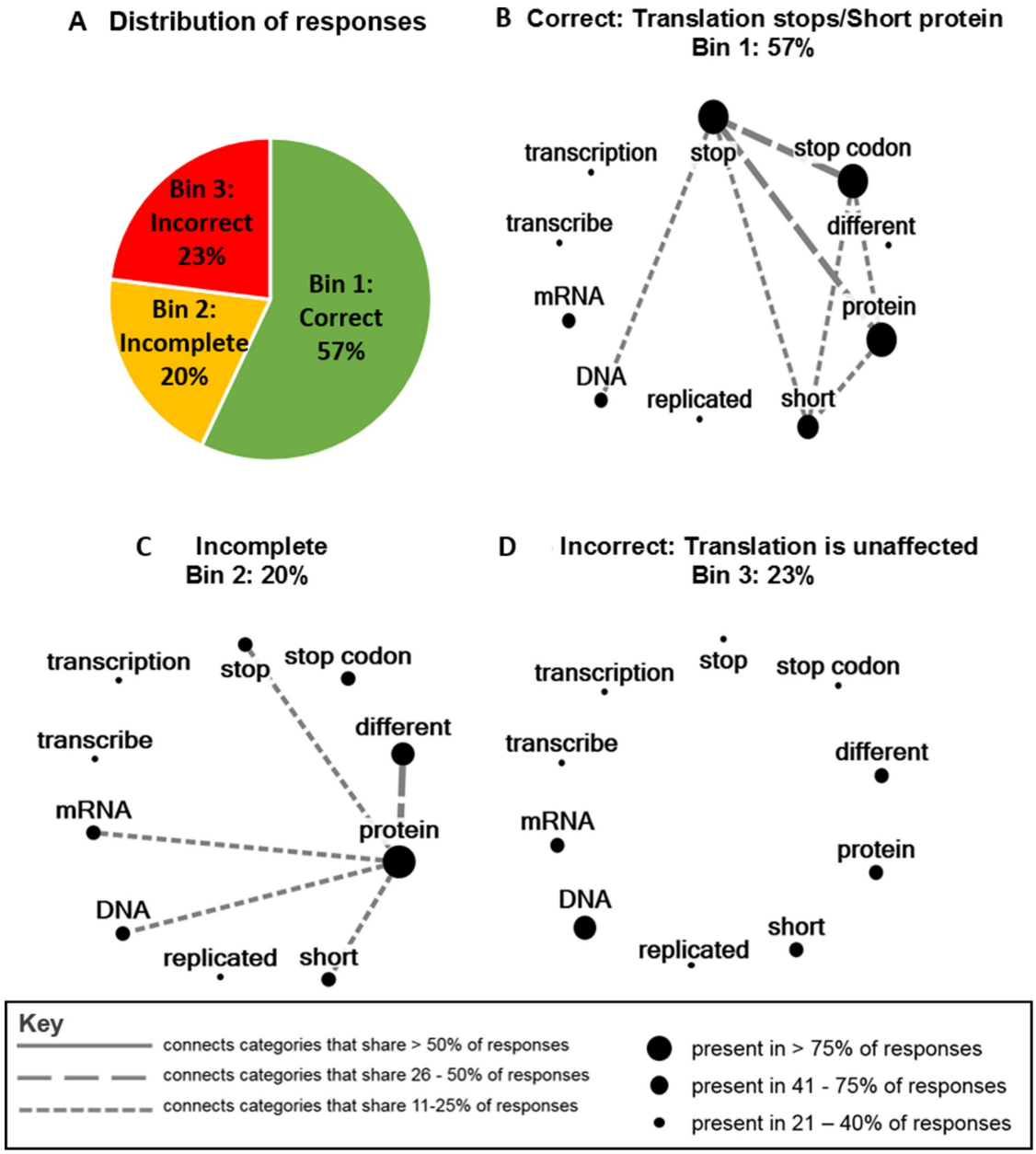

FIGURE 5. Student performance (A) and idea maps (B–D) for translation. Idea maps illustrate the frequency and cooccurrence of concepts in student writing.

Twenty-one percent of students also answered translation correctly but provided a mixture of incomplete/irrelevant and incorrect answers for replication and/or transcription (bins 2 and/or 3). These students are represented by multiple paths shown as thick black arrows in Figure 6, all ending in a green circle for translation.

For one of these paths, a student answered replication incorrectly, transcription correctly, and translation correctly (Figure 6C, bin 3 → bin 1 → bin 1): "The DNA strand will be too short due to the stop codon, causing huge deleterious mutations and likely a null allele. Replication will occur but stop at the new stop codon. >> It will not influence transcription. >> Translation will stop at the STOP codon instead of continuing normally."

In another path, students answered replication correctly, transcription incorrectly, but still answered translation correctly (Figure 6A, bin 1 → bin 3 → bin 1): "This will not alter DNA replication. >> Transcription will stop when the stop codon is reached. >> Proteins will only be made up until the stop codon."

Some students provided an incomplete or irrelevant response to the replication question, then answered transcription incorrectly, yet arrived at a correct answer for translation (Figure 6B, bin 2 → bin 3 → bin 1): "This is a nonsense mutation which will usually render the protein to be nonfunctional by stopping the reading sequence before transcription is complete. >> [T]ranscription will probably be stopped at the stop codon. >> [I]t won't occur for the remaining amino acids after the stop codon."

Downloaded from http://www.lifescied.org/ by guest on February 15, 2018

TABLE 6. Conceptual difficulties identified in student responses

Process	Conceptual difficulties
Replication	Stop codon stops replication, resulting in a shorter DNA strand. A shorter DNA strand results in a shortened mRNA strand and a shorten protein.
Transcription	Stop codon stops transcription and results in a shorter mRNA strand.
Translation	Translation does not take place when a stop codon is present. Translation does not stop when a stop codon is present, yet a shorter protein is produced. No protein would be produced. Fewer proteins would be produced during translation when a stop codon is present.

DISCUSSION

Lexical Analysis Can Successfully Group Diverse Responses That Exhibit Similar Levels of Correctness

In this study, student responses to questions posed about the effects of a mutation on replication, transcription, and translation were classified as correct, incomplete/irrelevant, or incorrect by both human scorers and through lexical and statistical analyses. As has been shown for questions in evolution (Ha *et al.*, 2011; Nehm and Haertig, 2012) and metabolism (Haudek *et al.*, 2012; Weston *et al.*, 2015), we have demonstrated that computerized lexical analysis can be used to identify student scientific and nonscientific ideas about the central dogma. Lexical and regression analyses successfully classified student responses, showing agreement with human coding (Tables 4 and 5). The majority of student responses were classified as correct or incorrect, with a small percentage of responses (7–20%) classified as incomplete or irrelevant (bin 2). This small group of incomplete/irrelevant responses (10%) was also observed in a previous study using lexical analysis to analyze student writing on acid–base chemistry (Haudek *et al.*, 2012). We were also able to trace student thinking across concepts, revealing connections between their ideas about all three processes. This analysis showed that ~30% of students had conceptual difficulties related to replication and transcription that would not have been revealed had they only been asked about translation (Figure 6). During learning, scientific and nonscientific ideas that students hold compete for use, with scientific ideas becoming more frequently and consistently used as expertise develops (Chi *et al.*, 1981; Opfer *et al.*, 2012). We conclude that constructed-response assessment coupled with lexical analysis can be a useful tool for capturing students' mixed mental models and evaluating how these models change over time.

Potentially Hidden Conceptual Difficulties and Their Causes

While student responses clearly displayed the previously identified conceptual difficulty that a stop codon causes a shorter mRNA strand to be produced (Smith *et al.*, 2008), we were able to elaborate on student ideas regarding these phenomena. When responses to each question were examined individually, more than one-third of students answered the translation question correctly. However, when we analyzed responses across all questions, we observed that an additional 30% of students arrived at correct responses to translation despite incorrect ideas regarding replication and transcription. This outcome demonstrates that conceptual difficulties about the central dogma can be hidden if students are asked to describe only one process or stage. We discuss below how these and other conceptual difficulties shown in Table 6 can be linked to the term "stop codon."

Some of student difficulties observed within our data set may come from the common understanding of the word "stop" compared with the meaning of the phrase "stop codon." While an expert would describe the role of the stop codon only in terminating translation, students often wrote that the stop codon causes replication and/or transcription to stop, producing shorter DNA and mRNA strands. Additionally, some students also appear to think that a stop codon will prevent the processes of transcription and translation from occurring; that is, these processes will not even begin. Incorrect conceptual understanding due to word association has been previously observed in students' responses to genetics and acid–base chemistry (Fisher, 1985; Haudek *et al.*, 2012). Students may create incorrect relationships between words that are frequently used in other science contexts. For example, students may incorrectly associate amino functional groups as having strongly acidic properties because of the cooccurrence of "amino" and "acid" in the term "amino acid" (Haudek *et al.*, 2012). In our study, stop codons can be deduced from the DNA and RNA sequences. Thus, students may erroneously associate stop codon, and thus stopping, with DNA and replication and/or RNA and transcription, as well as translation.

Because the correct product of translation in the presence of the introduced mutation is a shortened polypeptide chain, students who think that the DNA or mRNA is shortened are able to correctly answer a question about translation through an incorrect thought process. Therefore, students in this group would not experience any cognitive dissonance, as the outcome of a shortened polypeptide occurs despite their mistakes in thinking about replication and transcription. Accordingly, it is difficult to determine whether students have a complete understanding of the effect of a stop codon when students are only asked about the end product of translation, because their incorrect reasoning may be hidden. Therefore, to determine students' understanding of the processes of the central dogma, it is necessary to have them describe each of the processes individually.

To help students understand that the stop codon has a role only in terminating translation, we suggest that instructors change the way they refer to stop codons. Instructors can assist students with learning new vocabulary like "stop codon" by emphasizing 1) when the stop codon plays an active role (in translation) and 2) when the stop codon or precursors to the stop codon are inactive (during replication and transcription). Currently, the authors are collaborating with faculty at multiple institutions to develop an in-class activity to address students' difficulty understanding the role of the stop codon. In this activity, two points are emphasized to help student thinking: 1) the polymerases involved in replication and transcription read only one base at a time, and 2) the ribosome recognizes triplets (not single bases) during translation. Another way to prevent possible incorrect misinterpretations is to only use the phrase "translational stop codon" in the classroom. This more precise term could minimize potential ambiguity and make the association between stop codon and translation explicit.

Downloaded from http://www.lifescied.org/ by guest on February 15, 2018

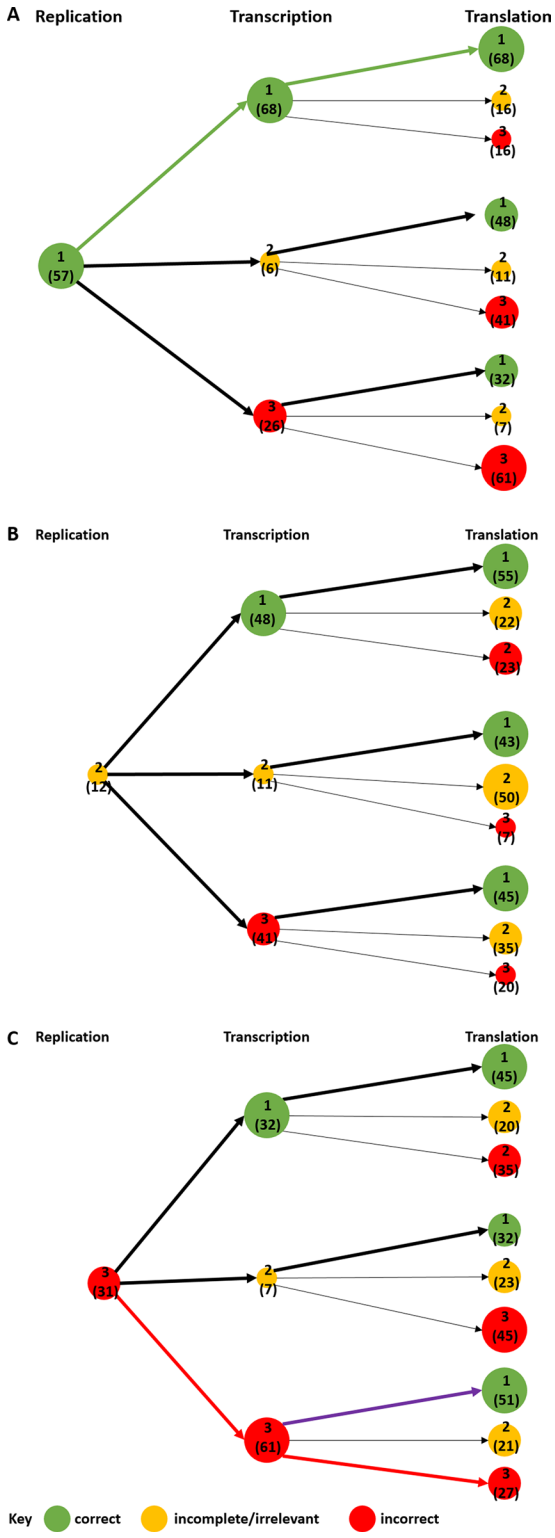

FIGURE 6. Relationships among student responses to each of the three questions. Correct (bin 1) responses are represented by green circles, incomplete/irrelevant responses (bin 2) by yellow circles, and incorrect responses (bin 3) by red circles. The size of circle represents the frequency of responses assigned to the bin. Numbers in parenthesis in a circle represent the percentage of responses in the bin. Numbers on the arrows represent the percentage of students who move from one bin to another. (A) Pathways for students who answered replication correctly. Students who answered all three questions correctly are represented by the green circles and arrows. (B) Pathways for students who answered replication incompletely/irrelevantly. (C) Pathways for students who answered replication incorrectly. Students who answered all three questions incorrectly are represented by the red circles and arrows. The purple arrow represents the students who got transcription and replication wrong but answered the translation question correctly. The thick black arrows (shown in all three panels) represent other pathways taken by students who arrived at correct answer for translation but show errors in their answers to replication and/or transcription.

Finally, to address potentially hidden conceptual difficulties, instructors may need to make changes in how they assess student understanding of the role of a stop codon within the central dogma processes. Assessments on the effects of stop codon mutations typically ask about just translation. Asking about how a mutation that produces a stop codon affects each of the central dogma processes can more fully reveal a student's sequence of thinking. For example, if students are not prompted to address each stage of a set of processes, they may omit information in their responses, because 1) they think this information is not relevant to the question, or 2) they cannot explain that stage or process. By explicitly asking students to write about each process, instructors can more easily diagnose the source of student misunderstanding.

Classroom Administration of the Question Set

We have shown that computer-assisted lexical analysis can be successfully used to identify student thinking from constructed-response assessments. This approach can be used in large-enrollment courses in which grading written assessments is too time-consuming. The analysis of student responses can then be used to provide feedback to both students and instructors about student thinking. There are a number of ways the questions can be used, including immediately after instruction to identify whether students are still struggling with these concepts and immediately before and after an activity specifically designed to target these concepts. In either use, student misunderstandings can be diagnosed and reported back to the students to further assist their learning in the spirit of just-in-time teaching (Novak *et al.*, 1999; Prevost *et al.*, 2013). For example, an instructor may administer these questions as homework after a module on the central dogma. The instructor may then use examples of student responses from each bin (correct, incomplete/irrelevant, and incorrect) to create a clicker question that prompts classroom discussion. Using students' own words as distractors can allow students to confront their incorrect understanding during class discussion. Alternatively, instructors may choose to use this feedback for more comprehensive changes, such as redesigning their lessons, activities, or homework, to be implemented the next time the course is taught.

Instructors interested in using these suites of questions, custom libraries, and scoring models for automated analysis of student responses may visit the Automated Analysis of Constructed Response (AACR) research group website at www.msu.edu\~aacr or contact the corresponding author.

Downloaded from http://www.lifescied.org/ by guest on February 15, 2018

ACKNOWLEDGMENTS

We thank members of the AACR research group for their help with data analysis and their thoughtful insights on the manuscript. We also thank Paula Lemons, Karen Pelletreau, Kate Semsar, and two anonymous reviewers for their feedback on this article. This work was funded through grants from the National Science Foundation (NSF; DUE 1022653, 1323022, 1322851, 1347578, and 1347626). Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC.

Birenbaum M, Tatsuoka KK (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. Appl Psychol Meas 11, 385–395.

Bowling BV, Acra EE, Wang L, Myers MF, Dean GE, Markle GC, Moskalik CL, Huether CA (2008). Development and evaluation of a genetics literacy assessment instrument for undergraduates. Genetics 178, 15–22.

Chi MTH, Feltovich PJ, Glaser R (1981). Categorization and representation of physics problems by experts and novices. Cogn Sci 5, 121–152.

Crocker L, Algina J (1986). Introduction to Classical and Modern Test Theory, Orlando, FL: Harcourt Brace Jovanovich.

Cronbach LJ (1984). Essentials of Psychological Testing, New York: Harper & Row.

Fisher KM (1985). A misconception in biology: amino acids and translation. J Res Sci Teach 22, 53–62.

Garvin-Doxas K, Klymkowsky MW (2008). Understanding randomness and its impact on student learning: lessons learned from building the biology concept inventory (BCI). CBE Life Sci Educ 7, 227–233.

Ha M, Nehm RH, Urban-Lurain M, Merrill JE (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations. CBE Life Sci Educ 10, 379–393.

Haudek KC, Prevost LB, Moscarella RA, Merrill J, Urban-Lurain M (2012). What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. CBE Life Sci Educ 11, 283–293.

IBM (2011a). IBM SPSS Modeler 14.2, IBM Corporation.

IBM (2011b). IBM SPSS Text Analytics for Surveys 4.0.1 User's Guide, Chicago, IL.

Kalas P, O'Neill A, Pollock C, Birol G (2013). Development of a meiosis concept inventory. CBE Life Sci Educ 12, 655–664.

Kaplan JJ, Haudek KC, Ha M, Rogness N, Fisher DG (2014). Using lexical analysis software to assess student writing in statistics. Technology Innovations in Statistics Education 8. http://escholarship.org/uc/item/57r90703 (accessed 5 February 2015).

Keys CW (1999). Revitalizing instruction in scientific genres: connecting knowledge production with writing to learn in science. Sci Educ 83, 115–130.

Kuechler WL, Simkin MG (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. Decis Sci J Innov Educ 8, 55–73.

Lemke JL (1990). Talking Science: Language, Learning, and Values, Norwood, NJ: Ablex.

Lewis J, Leach J, Wood-Robinson C (2000). All in the genes? Young people's understanding of the nature of genes. J Biol Educ 34, 74–79.

Marbach-Ad G (2001). Attempting to break the code in student comprehension of genetic concepts. J Biol Educ 35, 183–189.

Martinez ME (1999). Cognition and the question of test item format. Educ Psychol 34, 207–218.

Menard S (2002). Applied Logistic Regression Analysis, Thousand Oaks, CA: Sage.

Mills Shaw KR, Van Horne K, Zhang H, Boughman J (2008). Essay contest reveals misconceptions of high school students in genetics content. Genetics 178, 1157–1168.

Nehm RH, Haertig H (2012). Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. J Sci Educ Technol 21, 56–73.

NGSS Lead States (2013). Next Generation Science Standards: For states, by States, Washington, DC: National Academies Press.

Novak GM, Patterson ET, Gavrin AD, Christian W, Forinash K (1999). Just in time teaching. Am J Phys 67, 937.

Opfer JE, Nehm RH, Ha M (2012). Cognitive foundations for science assessment design: knowing what students know about evolution. J Res Sci Teach 49, 744–777.

Peng C-YJ, Lee KL, Ingersoll GM (2002). An introduction to logistic regression analysis and reporting. J Educ Res 96, 3–14.

Prevost LB, Haudek KC, Norton Henry E, Berry MC, Urban-Lurain M (2013). Automated text analysis facilitates using written formative assessments for just-in-time teaching in large enrollment courses. Paper presented at 2013 ASEE Annual Conference & Exposition, Atlanta, GA. https://peer.asee.org/19250.

Price RM, Andrews TC, McElhinny TL, Mead LS, Abraham JK, Thanukos A, Perez KE (2014). The Genetic Drift Inventory: a tool for measuring what advanced undergraduates have mastered about genetic drift. CBE Life Sci Educ 13, 65–67.

Shrout P, Fleiss J (1979). Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86, 420–428.

Smith MK, Knight JK (2012). Using the Genetics Concept Assessment to document persistent conceptual difficulties in undergraduate genetics courses. Genetics 191, 21–32.

Smith MK, Wood WB, Knight JK (2008). The Genetics Concept Assessment: a new concept inventory for gauging student understanding of genetics. CBE Life Sci Educ 7, 422–430.

Vosniadou S, Brewer WF (1992). Mental models of the earth: a study of conceptual change in childhood. Cogn Psychol 24, 535–585.

Weston M, Haudek KC, Prevost L, Urban-Lurain M, Merrill J (2015). Examining the impact of question surface features on students' answers to constructed-response questions on photosynthesis. CBE Life Sci Educ 14, ar19.

Wood-Robinson C (2000). Young people's understanding of the nature of genetic information in the cells of an organism. J Biol Educ 35, 29–36.

Wright LK, Fisk JN, Newman DL (2014). DNA → RNA: what do students think the arrow means? CBE Life Sci Educ 13, 338–348.

Downloaded from http://www.lifescied.org/ by guest on February 15, 2018