



EAB

Text Parsing and Handling Strings in Construct

A Rapid Insight Deep Dive

March 10, 2022



Many data projects hit a real slowdown when it comes to a need for processing text data. Multi-key fields often need to be parsed, code values might need to be translated, manual entry might have introduced typos, or maybe you only need the first letter out of the whole string (another word for text data). The fix might seem easier to do by hand but scaling up a manual fix to include thousands of rows isn't feasible. Even so, it may seem insurmountable to create a systematic process to parse, clean, substitute, or otherwise transform the text you need to analyze.

In this webinar, expect to learn how Construct can:

- Parse out multi-key fields (e.g., "AARN, 20FA, ")
- Replace values within strings at multiple levels of precision
- Account for typos in manual entry fields (e.g., "main street" vs. "main st.")
- Shorten or lengthen values (e.g., adding or removing leading zeroes)
- Extract subfields from strings dynamically

This webinar will help you leverage Construct towards scalable text transformations on your data. We're not assuming any background experience with text manipulations, so come as you are, and find out how you can expertly manipulate strings.



Meet Your Presenters



James Cousins
Edify Product Analyst



Lily Brennan
Edify Product Analyst

Submit a Question or Comment

The screenshot displays the EAB Virtual Session interface. At the top left is the EAB logo, consisting of a blue circle with a white building icon and the text "EAB". Below the logo, the text "EAB Virtual Session" is displayed in white and teal. Two callout boxes with orange borders provide instructions: "Use the Chat feature to send messages to all panelists or everyone" and "Use the Q&A feature to ask questions". At the bottom, a dark navigation bar contains "Audio Settings" with an upward arrow, "Chat" with a speech bubble icon, "Q&A" with a speech bubble icon, and "Leave Meeting" in red text.

EAB

EAB
Virtual Session

Use the Chat feature to send messages to all panelists or everyone

Use the Q&A feature to ask questions

Audio Settings ^

Chat

Q&A

Leave Meeting

Turn on Captions

The image shows a Zoom meeting interface with a white header bar at the top containing the EAB logo and the text "EAB". Below the header is a dark blue background with a pattern of overlapping circles and lines. In the center, the text "EAB Virtual" is visible. A yellow-bordered callout box with a black border contains the text "Enable an automated Live Transcript – Show Subtitle or View Full Transcript". A yellow line points from this box to a dark grey menu that is open, showing three options: "Show Subtitle" (highlighted in blue), "View Full Transcript", and "Subtitle Settings...". At the bottom of the screen, there is a dark grey control bar. On the left, it says "Audio Settings" with an upward arrow. In the center, there is a "CC" icon above the text "Live Transcript". On the right, there is a red "Leave Meeting" button.

EAB

EAB
Virtual

Enable an automated Live Transcript –
Show Subtitle or View Full Transcript

- Show Subtitle
- View Full Transcript
- Subtitle Settings...

Audio Settings ^

CC
Live Transcript

Leave Meeting

- 1 **Rapid Insight and EAB**
- 2 Overview of the Presentation Structure
- 3 Text Parsing Challenges
- 4 Audience Q & A

A New Partnership to Accelerate Data Democratization in Higher Education

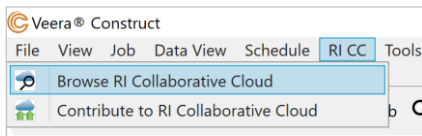


Learn More:

Read more about EAB's latest partnership at
<https://eab.com/rapidinsight/>

- 1 Rapid Insight and EAB
- 2 **Overview of the Presentation Structure**
- 3 Text Parsing Challenges
- 4 Audience Q & A

Rapid Insight Collaborative Cloud



Rapid Insight® Collaborative Cloud

(All) Search

Filter By Contributor: (All) Sort By: Contribution Date

Deep Dive Webinar: Text Parsing and Handling Strings in Construct by James Cousins EAB This job represents simple versions of 5 distinct types of text parsing challenges that a user might face while using Construct. The "Embed Data" nodes contain the data that each challenge requires. Challenge #1: Parsing Composite Keys ("Multi-Key Fields") Challenge #2: Replacing Entire or Partial Strings	Date Posted: 3/8/2022 Times downloaded: 0 # Comments: 0
IPEDs Spring Collection Import Specification Schema File Generation by Lily Brennan EAB This job facilitates IPEDs import formatting by creating the "schema" file that enforces the column position and length that each survey requires. Each "Embed Data" node runs to a .txt output on your machine.	Date Posted: 3/4/2022 Times downloaded: 0 # Comments: 0
DTSC Admin Services Satisfaction Survey Analysis (2) by Robertson, Aaron@DTSC DTSC DTSC Admin Satisfaction Data	Date Posted: 1/15/2022 Times downloaded: 2 # Comments: 0
Cartesian Joins Example - Adding Non-Enrollment Terms by Jon MacMillan RAPIDINSIGHTINC	Date Posted: 9/23/2021 Times downloaded: 11

Page 3 of 3 Loaded



Details

Title: Deep Dive Webinar: Text Parsing and Handling Strings in Construct

Contributor: James Cousins

Date Posted: 3/8/2022 11:41:00 AM Category: Job

Description:

This job represents simple versions of 5 distinct types of text parsing challenges that a user might face while using Construct. The "Embed Data" nodes contain the data that each challenge requires.

Challenge #1: Parsing Composite Keys ("Multi-Key Fields")
Challenge #2: Replacing Entire or Partial Strings
Challenge #3: Accounting for Common Typos in Manual Entry Fields
Challenge #4: Padding Truncated Fields to Fit Desired Length
Challenge #5: Extracting Sub-Strings Dynamically

History:

3/8/2022 11:41 AM - Added Attachment "Deep Dive Webinar: Text Parsing and Handling Strings in Construct.vcj"

Attachments (Optional)

+ Add File Attachment X Remove File Attachment

I accept the [Terms & Conditions](#)

Comments Delete Download



Classifying the Challenge

It helps to recognize the type of challenge so that you know where to begin solving it.



Common Solution

While there's typically not one singular solution, there's usually a useful go-to technique.



Live Example

We'll show the solution in action, highlighting several alternative solutions along the way.

- 1 Rapid Insight and EAB
- 2 Overview of the Presentation Structure
- 3 **Text Parsing Challenges**
- 4 Audience Q & A

Poll

How would you rate your proficiency using Construct to handle text data?

- a) I have never used Construct to work with text data
- b) I have used Construct to work with text data, but I would like to learn more best practices
- c) I am confident handling text data in Construct



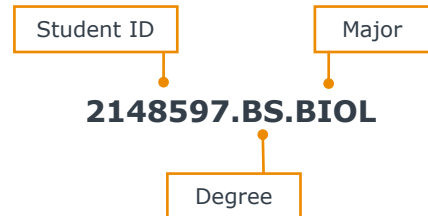
- 1 Parsing multi-key fields (e.g., "2148597.BS.BIOL")
- 2 Replacing targeted text at multiple levels of precision
- 3 Accounting for typos in manual entry fields
- 4 Shortening (trimming) or lengthening (padding) values
- 5 Extracting subfields from strings dynamically

Challenge 1

Parsing multi-key fields
(e.g., "2148597.BS.BIOL")

▶ Multi-Key Fields

- Technically referred to as a Composite Key
- Concatenates multiple characteristics as a singular key for the record





Useful Strategies

- Other Columns in the same table
 - Composite keys are usually comprised of information stored in the same table, across multiple columns
- SUBFIELD()
- SUBSTRING()
- LEFT(), potentially combined with REMOVELEFT()
 - See also, RIGHT(), REMOVERIGHT()

2148597.BS.BIOL

Is the first field always 7 characters long?

Is the delimiter always a "."?

Is the third portion of the key always the Major or only when a degree is available?



Multi-Key Field Example

What Transform Formulas Would Get You What You Need?

Original Field	Student ID	Degree	Major
2148597.BS.BIOL	2148597	BS	BIOL

Parsed Variable	Using LEFT()/RIGHT() and REMOVELEFT()/REMOVEDRIGHT()	Using SUBFIELD()	Using SUBSTRING()
Student ID	LEFT ([A],7)	SUBFIELD ([A],1,',')	SUBSTRING ([A], 1, 7)
Degree	LEFT (REMOVELEFT ([A],8),2)	SUBFIELD ([A],2,',')	SUBSTRING ([A], 9, 2)
Major	LEFT (REMOVELEFT ([A],11),4)	SUBFIELD ([A],3,',')	SUBSTRING ([A], 12, 4)

Challenge 2

Replacing targeted text at
multiple levels of precision

Take What You Like, Leave What You Don't



Program
Automotive Technology – AS
Marine Biology – BS
Religious Studies - MA

Text Replacements

- The idea is that **at least some** of the text needs to be replaced
 - If the entire string needs replacing, it is just a Text Replacement
 - If just some portion of the string needs replacing, it is a “Sub-String” Replacement

Partial Replacement

Cleanse Node

Entire Replacement



"Replace Sub-String"

Replace Modify

Trim
 Obfuscate
 Sub-String Replacement

Replace all occurrences of:

With:

Add

Sub-String Replacements:

↑ ↓ ✕

'#' -> '#'
'$' -> '\$'
'%' -> '%'

Name: Add Clear

"Replace Text"

Replace Modify

When

TextField

Replace with:

Replace With Null

Name: Add Clear

Example: Merging Degree Data

Program
Automotive Technology – AS
Marine Biology – BS
Religious Studies – MA
Psychology - BS

Program
Automotive Technology – A.S.
Marine Biology – B.S.
Religious Studies – M.A.
Psychology – B.S.

- Replacing the entire string in this case would work, but be sub-optimal
- To replace “Marine Biology – BS” and “Psychology – BS” with “Marine Biology – B.S.” and “Psychology – B.S.”, you’d need an operation for each *entire string*.
- Instead, use “Replace Sub-String” to just replace all instances of “BS” with “B.S.”, “AS” with “A.S.”, and so forth

Challenge 3

Accounting for typos in
manual entry fields

How Typos End Up in Data



Two Conditions Must Be Met

1

2



Manual Entry Fields

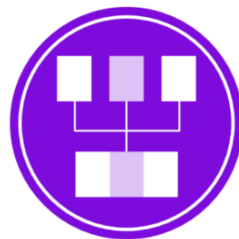
Naturally, *unexpected* typos only occur when data entry occurs through manual entry. Surveys and administrative data entry without validation are frequent causes.

A Known Set of Possibilities

You can't systematically diagnose text as having typos without a known set of "correct" values. These valid values become the foundation of corrective rules in your data preparation.

Local Solutions:

- Use a Cleanse node and build replacement rules
 - For names, departments, and other entities with official names, use text replacements
 - For acronyms, a sub-string replacement may be worth considering
- Merging with a Lookup Table is a more robust resource that can be modified and maintained easily



Systematic Solution:

- If possible, reduce the number of manual entry fields
- Introduce validation rules to prevent typos from cascading into operational data

Example: A Manually Entered "Role"

What is your role at the institution?

Student

Student

Studnet

Faculty

Daculty

Stfaf

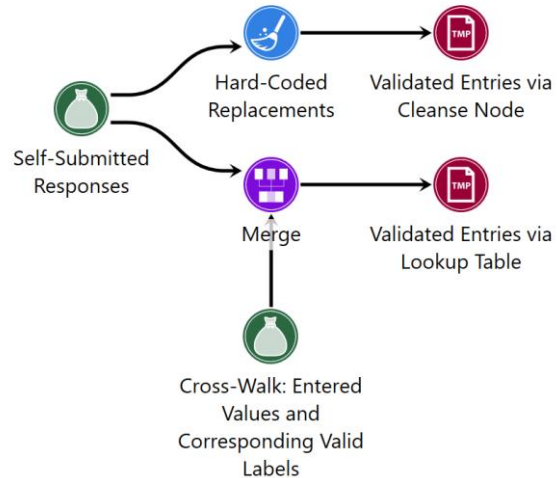
Staff

Known Valid Entries

Student

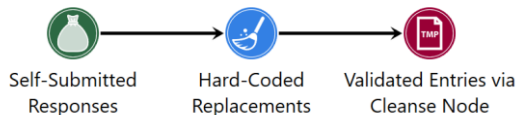
Faculty

Staff



Solving the Issue Using the Cleanse

Manual Entry	Validated
Student	Student
Student	Student
Studnet	Student
Faculty	Faculty
Daculty	Faculty
Stfaf	Staff
Staff	Staff



Hard-Coded Replacements: [Cleanse]

File Edit Sort Help

Replace Modify

When: What is your role at the institution?

= Daculty

Replace with: Faculty

Replace With Null

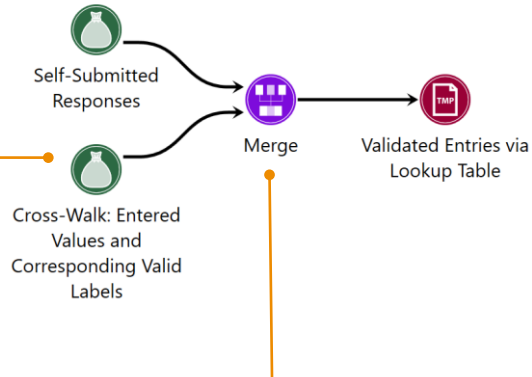
Name: What_is_your_role_at_the_inst Update Clear Sample rows: 5000

Rank #	Cleanse Operation Nam	Cleanse Type	Column Name	Cleanse Operation
1	What_is_your_role_at_the	Replace Text	What is your role	Replace 'Daculty' with 'Faculty'
2	What_is_your_role_at_the	Replace Text	What is your role	Replace 'Stfaf' with 'Staff'
3	What_is_your_role_at_the	Replace Text	What is your role	Replace 'Studnet' with 'Student'

Ready.

Solving the Issue Using a Merge and a Crosswalk

Entered Value	Validated Label
Student	Student
Student	Student
Studnt	Student
Stundet	Student
Faculty	Faculty
Daculty	Faculty
Favulty	Faculty
Facluty	Faculty
Staff	Staff
Stff	Staff
Staf	Staff
Stfaf	Staff



• Merge: [Merge]

File Edit Sort Help

X [Icons]

(1) Self-Submitted Responses ^

*
What is your role at the institut

(2) Cross-Walk: Entered Values ar ^

*
Entered Value
Validated Label

1) "Crosswalk" is a common term for a "reference table", "lookup table", and perhaps other terms. It provides a mapping from one value to a corresponding one, as in the case of a numeric code corresponding to a department name.

Poll

Which approach would you choose?

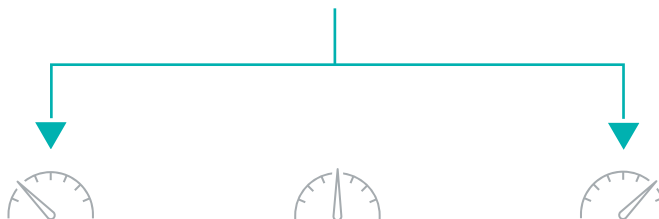
- a) Use the Cleanse node to find and replace values
- b) Use a Merge node to cross reference a lookup table



Challenge 4

Shortening (trimming) or
lengthening (padding) values

Values that are not the right length



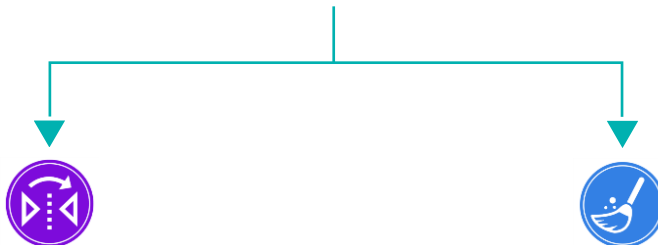
Too Short

9 out of 10 times, this means that leading zeroes got dropped by some data-read or formatting process

Too Long

Either because A) only the first portion is valuable, or B) because the value is formatted to include blank characters ("whitespace") out to a certain length

Fixing values that are not the right length



Too Short: Transform

The Transform's "PADLEFT()" and "PADRIGHT()" functions allow users to lengthen a field and impute specific values in case the original text is not long enough.

Too Long: Cleanse

The "trim" functionality in a Cleanse node removes all leading and trailing whitespace in a field.



Interestingly, the Transform *does* have the "TRIM()" and the "LEFT()" functions, which can help with trimming down fields too.

Padding Truncated Zip Codes

Obs #	City	State	Zip Code		5 Digit ZipCode
1	Madison	NH	3849	→	03849
2	Madison	WI	53558	→	53558
3	Madison	CT	6443	→	06443



Binning Multi-Variable Formula Text Functions Date/Time

Assign Variables: [A] - Zip Code

Enter a formula: (example A+B-)

PADLEFT([A],5,'0')

Function Definition:

Result Type: Text

New Variable Name: 5 Digit ZipCode

Update Cancel Delete

Challenge 5

Extracting subfields from
strings dynamically

▶ Subfields not Conforming to Constant “Length” or “Order” Assumptions

- This general requirement can get *very* complex in some cases
 - These are cases where you can’t count on the target sub-string being the “first”, or any specific position as a rule
 - You also can’t count on the target sub-string beginning at a certain character position (e.g., “the 8th through 14th characters”)
- You’ll need to find some pattern, even if it’s a complex one

47419,50544,0
67306,69092
47419, 80479, 22583, 129084, 60802

This is an excellent case, where just the highlighted element should be returned from each row’s value.

Go-To Tools for Dynamic Subfields



Functions	Description	Frequency of Use
SUBFIELD()	Returns the n^{th} subfield from within a string as delimited by a specified character.	Common
SUBFIELDCOUNT()	Returns the number of subfields found within a string as delimited by a specified character.	Uncommon
CHARINDEX()	Returns the starting position of a specified expression within a string- and a "0" if not found.	Uncommon
MATCHESREGEX()	Allows the user to stipulate a "regular expression" and tests whether the string matches that set of rules.	Rare
LEFT()/RIGHT()	Returns the specified number of characters, starting from either the left or the right.	Common
REMOVELEFT()/REMOVERIGHT()	Removes the specified number of characters from a string, starting from either the left or the right.	Uncommon

Example: Returning the Last, Non-Zero Code

Target Outcome:

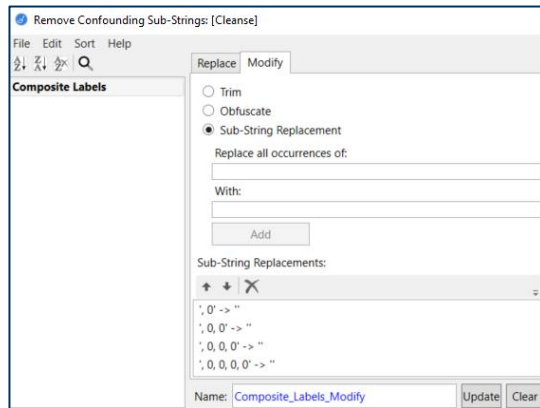
Systematically creating the column on the right



Obs #	Composite Labels	Last Label in String
1	47419, 50544, 0	50544
2	67306, 69092	69092
3	67306, 69092	69092
4	67306, 69092	69092
5	67306, 69092	69092
6	67306, 69092	69092
7	67306, 69092	69092
8	67306, 69092	69092
9	67306, 69092	69092
10	35977, 69092, 0	69092
11	35977, 69092, 0	69092
12	35977, 69092, 0	69092
13	35977, 69092, 0	69092
14	35977, 69092, 0	69092
15	35977, 69092, 0	69092
16	35977, 69092, 0	69092
17	47419, 80479, 22583, 129084, 50186, 46092, 60802	60802
18	47419, 80479, 22583, 129084, 50186, 46092, 60802	60802
19	61387, 0, 0, 0	61387
20	61387, 0, 0, 0, 0	61387

Step 1

Use a sub-string replacement operation to eliminate the undesirable values from each string



Example Continued

Returning the Last, Non-Zero Code

Target Outcome:

Systematically creating the column on the right



Obs #	Composite Labels	Last Label in String
1	47419, 50544, 0	50544
2	67306, 69092	69092
3	67306, 69092	69092
4	67306, 69092	69092
5	67306, 69092	69092
6	67306, 69092	69092
7	67306, 69092	69092
8	67306, 69092	69092
9	67306, 69092	69092
10	35977, 69092, 0	69092
11	35977, 69092, 0	69092
12	35977, 69092, 0	69092
13	35977, 69092, 0	69092
14	35977, 69092, 0	69092
15	35977, 69092, 0	69092
16	35977, 69092, 0	69092
17	47419, 80479, 22583, 129084, 50186, 46092, 60802	60802
18	47419, 80479, 22583, 129084, 50186, 46092, 60802	60802
19	61387, 0, 0, 0	61387
20	61387, 0, 0, 0	61387

Step 2

- Use a Transform to count the number of subfields left, and return the last subfield from each value
- `SUBFIELD([A],SUBFIELDCOUNT([A],','),',')`
 - `SUBFIELDCOUNT([A],',')` returns the number of fields each row has
 - `SUBFIELD([A], ...,','')` uses that result to return the last subfield's text for each row

Binning Multi-Variable Formula Text Functions Date/Time

Assign Variables: [A] - Composite Labels

Enter a formula: (example A+B- P) **SUBFIELD([A],SUBFIELDCOUNT([A],','),',')**

Function Definition:

Result Type: Text

New Variable Name: Last Label in String

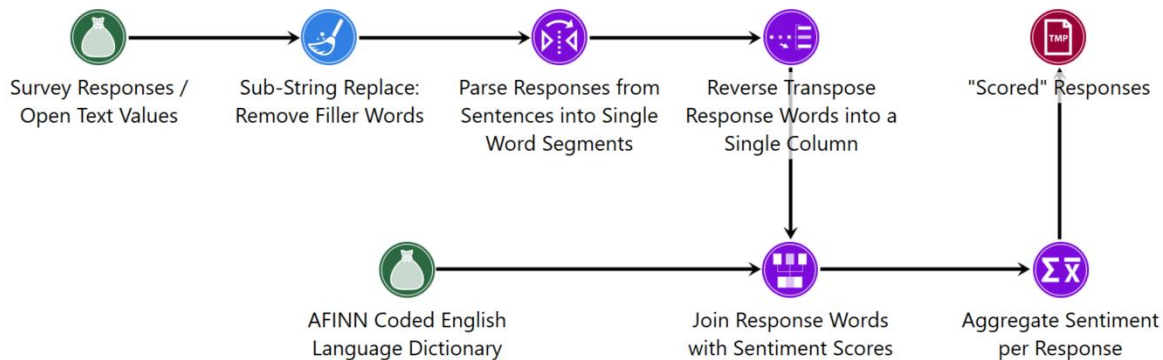
Update Cancel Delete

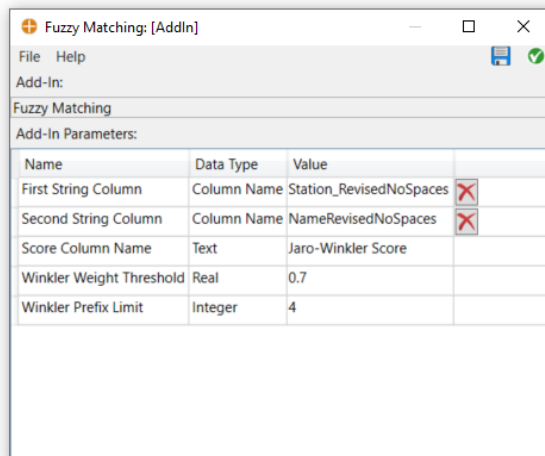
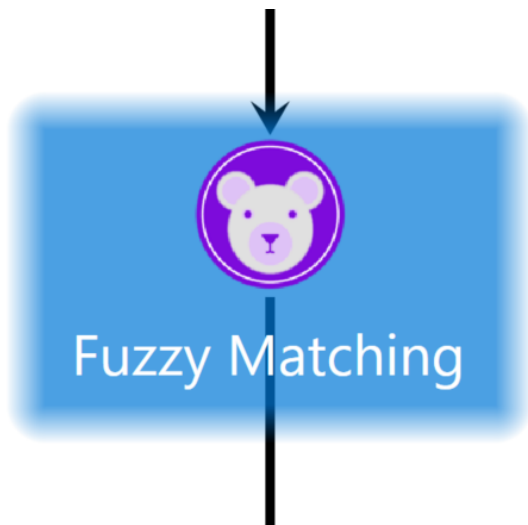
Bonus!



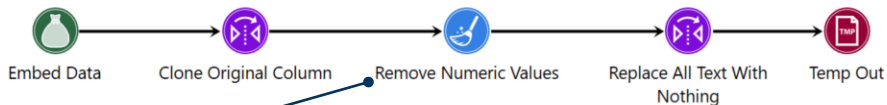
Advanced possibilities

Sentiment Analysis





Returning Numeric Fields Only



Replace Modify

Trim

Obfuscate

Sub-String Replacement

Replace all occurrences of:

With:

Add

Sub-String Replacements:

+	-
+	X
'0'	-> ''
'1'	-> ''
'2'	-> ''
'3'	-> ''
'4'	-> ''
'5'	-> ''
'6'	-> ''
'7'	-> ''
'8'	-> ''
'9'	-> ''

Name: Code_Modify Update Clear

Binning Multi-Variable Formula Text Functions Date/Time

Assign Variables: ↑ ↓

[A] - Code Original

[B] - Code

Enter a formula: (example A+B-)

REPLACE([A], [B], "")

Function Definition:

Result Type: Text

REPLACE() returns a text field where all occurrences of a specified expression are replaced with a new expression

Poll

Do you feel more comfortable with text data in Construct now?

- a) Yes, definitely more comfortable than before
- b) I feel the same level of confidence as before
- c) I am more confused now than I was before



- 1 Rapid Insight and EAB
- 2 Overview of the Presentation Structure
- 3 Text Parsing Challenges
- 4 Audience Q & A

Submit a Question or Comment



The screenshot displays the EAB Virtual Session interface. At the top left is the EAB logo, consisting of a blue circle with a white building icon and the text "EAB". Below the logo, the text "EAB Virtual Session" is displayed in white and teal. Two callout boxes with orange borders provide instructions: "Use the Chat feature to send messages to all panelists or everyone" and "Use the Q&A feature to ask questions". At the bottom, a dark blue navigation bar contains the following elements from left to right: "Audio Settings" with an upward arrow, "Chat" with a speech bubble icon, "Q&A" with a speech bubble icon, and "Leave Meeting" in red text.

EAB

EAB
Virtual Session

Use the Chat feature to send messages to all panelists or everyone

Use the Q&A feature to ask questions

Audio Settings ^

Chat

Q&A

Leave Meeting

Quick Poll

How was today's session?

Please take a few minutes to complete the survey to provide additional feedback! A link will be placed in the Chat and you'll receive a follow-up email.



Washington DC | Richmond | Birmingham | Minneapolis

202-747-1000 | eab.com

 [@eab](https://twitter.com/eab)  [@eab_](https://www.linkedin.com/company/eab_)  [@WeAreEAB](https://www.facebook.com/WeAreEAB)  [@eab.life](https://www.instagram.com/eab.life)

