12

The Data Profiler tests for valid and invalid values

- The Data Profiler tests
- ~~How testing results are stored in the EME~~
- Values that match more than one test
- Important assumptions about validity
- Invalidity tests
- Another definition of invalid
- ~~How a value can be both valid and invalid~~

The Data Profiler tests

This list describes the Data Profiler validity and invalidity tests, in the order they are performed. Note that the invalidity tests precede the validity tests.

If a value matches one ~~and only one, not more~~ of these tests, the test name (**Null**, **Invalid value**, **Invalid character**, etc.) appears in the profile results as the **REASON** in the **Data Quality Summary**, and as the **DESCRIPTION** in the Common Values reports.

If a value matches multiple tests, it appears in the profile results with one of the descriptions listed in "Values that match more than one test".

Test: NULL

If the record format defines NULL, the Data Profiler compares the Physical Element value to the NULL value. ~~If the value is NULL, that's how it's treated.~~ For more information, see "Interpreting a value as NULL".

Test: Invalid value

~~Data profiler does a comparison between~~ the Physical Element value ~~and~~ the Validation Spec ~~invalid values.~~ If it matches any invalid value, ~~it is treated as~~ *invalid*.

Test: Invalid ~~value~~ character

The Physical Element value is compared against the character set specified in the Validation Spec. If the value contains ~~any~~ characters that are not listed in **Value Must Only Use These Characters**, ~~it is treated as~~ *invalid*.

Test: Invalid lookup

The value is ~~looked up in~~ the invalid lookup file. If the value ~~is found there~~, it is ~~treated as~~ *invalid*.

Test: Invalid expression

The Physical Element value is ~~tested~~ against each of the invalid expressions in order. If the invalid expression returns NULL or an error, the value is treated like it did not trigger the expression. If invalid expressions return a non-zero result, the value is ~~treated as~~ *invalid*. 1 is the index of the first valid expression in the list.

Because these are all tests, is it necessary to prefix each test name with "Test."

Does a value match a test, or is the Data Profiler testing a value against a test and then returns the test name?

Consider: The Data Profiler tests for invalid and valid values. The tests described are in the order that they are performed.

Should test names be in bold?

Latin abbreviations OK?

If new headings are added, include mini TOC after the introductory information: • Tests for invalid values • Tests for valid values

Not in the doc

Why is Validation Spec capitalized? It seems like this is a generic term.

I'm not sure what's being tested — the Physical Element value or expressions? Does the Physical Element value trigger expressions?

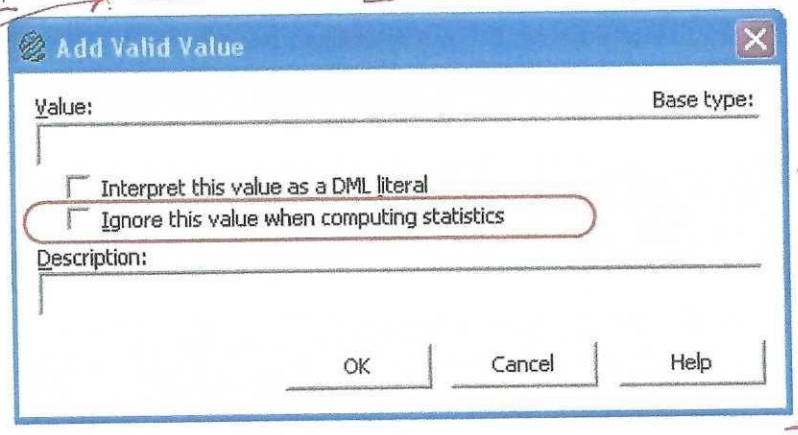
Not sure what this has to do with the test.

How is the expression triggered? Is the test comparing the Physical Element value to a list of invalid expressions?

Tests for valid values ← suggest a heading

Test: Ignore this value when computing statistics

The Physical Element value is compared against the list of valid values in the Validation Spec. If the value matches a valid value that is flagged as **Ignore this value when computing results**, the Physical Element value is considered *valid* but is not included in the calculation of the **Histogram**, Total values, Valid values, Invalid values, Mean value, Distinct values, Unique values, Minimum value, Maximum value, Empty values, Blank values, Normal values, Ascending pairs, Descending pairs, or **Cross-Field Relationships**. This setting is in the **Add valid values** dialog:



Test: is_valid

The Physical Element value is tested by calling the DML function **is_valid**. If is_valid returns **0** (**False**), the value is treated as *invalid*.

- **NOTE:** If the Physical Element DML definition contains the **is_valid_XXX** field validity function, the **is_valid_XXX** function is currently ignored.

Test: Valid value

The Physical Element value is looked at in comparison to the valid values in the Validation Spec. If the value matches *any* of these valid values, the value is treated as *valid*.

Test: Valid range

The Physical Element value is compared to the valid ranges in the Validation Spec. If the value falls within any of the ranges (including if it matches either endpoint of any range), the value is treated as *valid*.

Test: Valid pattern

The Physical Element value is compared to the valid patterns in the Validation Spec. If there's any match, it's treated as *valid*.

Test: Valid expression

The Physical Element value is tested against each of the valid expressions in order; 1 is the index of the first expression in the list. Values for expressions that return NULL or an error are treated like they did not trigger the expression. If expressions return a non-zero result, the value is ~~treated as~~ *valid*.

Test: Valid lookup

The value is ~~looked up in~~ the valid lookup file. If the value ~~found~~ in the valid lookup file, it is ~~treated as~~ *valid*.

Values that match more than one test

~~When~~ the Data Profiler tests ~~have~~ run on all the values in the dataset for a given Physical Element, the Data Profiler analyzes the results and decides how to group the statistics for values that matched more than one test. For matching values, the **DESCRIPTION** value in the **Data Quality Summary** or the **REASON** value in the Common Value reports may be one of the following (in addition to those sited in the Data Profiler tests):

- Valid
- Both valid and invalid
- Multiple reasons for invalidity
- Multiple reasons for validity
- Fails all definitions of validity

Valid

~~A description of~~ *valid* mean that the Physical Element value did not match any invalidity test, and there are no valid values, ranges, patterns, expressions, or lookups defined.

Both valid and invalid

One value of the Physical Element being profiled is *valid* in one record while the same value is *invalid* in another record. This can happen when one or more validity or invalidity expressions refer to a Physical Element in the record other than the Physical Element being tested.

The result can be one of these:

- One value in one record causes a validity expression to result in **True** while the same value in another record causes a validity expression to result in **False**.
- One value in one record causes an invalidity expression to result in **True** while the same value in another record causes an invalidity expression to result in **False**.

- One value in one record causes a *validity* expression to result in **True** while the same value in another record causes an *invalidity* expression to result in **True**.
- One value in one record causes an *invalidity* expression to result in **True** while the same value in another record causes an *invalidity* expression to result in **True**.

For more information, see "How a value can be both valid and invalid"

Multiple reasons for validity

Multiple instances of a Physical Element value triggers morethan one test for *validiity*. This can happen when multiple validity expressions refer to fields in teh record other than the field being tested.

The result can be that a value is valid in one record because it triggers the *first* validity expression in one record but triggers the *second* validity expression in another record.

Multiple reasons for invalidity

The meaning of ***multiple reason for invalidity*** depends on the setting of **Reasons for Invalidity** when the profile. For more information, see "The setting of Reasons for invalidity"

If **Report first reason only** was selected when when running a job, ***multiple reasons for invalidity*** means the value is *invalid* for one reason in one or more records, and invalid for a different reason in one or more other records. This can occur when an invalidity expression uses the value of another Physical Element in the record. For example, a value might trigger the third invalidity expression in one record but trigger the fourth invalidity expression in another record.

If **Report all reasons** was selected when running a job, ***multiple reasons for invalidity*** can means the same thing as was described in the preceding paragraph. Or, it can mean that the testing of a Physical Element value in one record had at least two of the following results:

- The value matches an invalid value.
- The value contained an invalid character.
- The value is in a lookup file of invalid values.
- The value triggered an expression that indicates invalidity.
- The value is invalid for the data type defined in the dataset's record format.
- The value fails all definitions of validity.

Fails all definitions of validity

Valid values, valid ranges, valid patterns, valid expressions, and/or valid lookups are defined, however, the value does not match any of them. And, the value passes the **is_valid** test and is not flagged **Ignore this Value when Computing Results**.

Important assumptions about validity

There are four important assumptions the Data Profiler makes about validity:

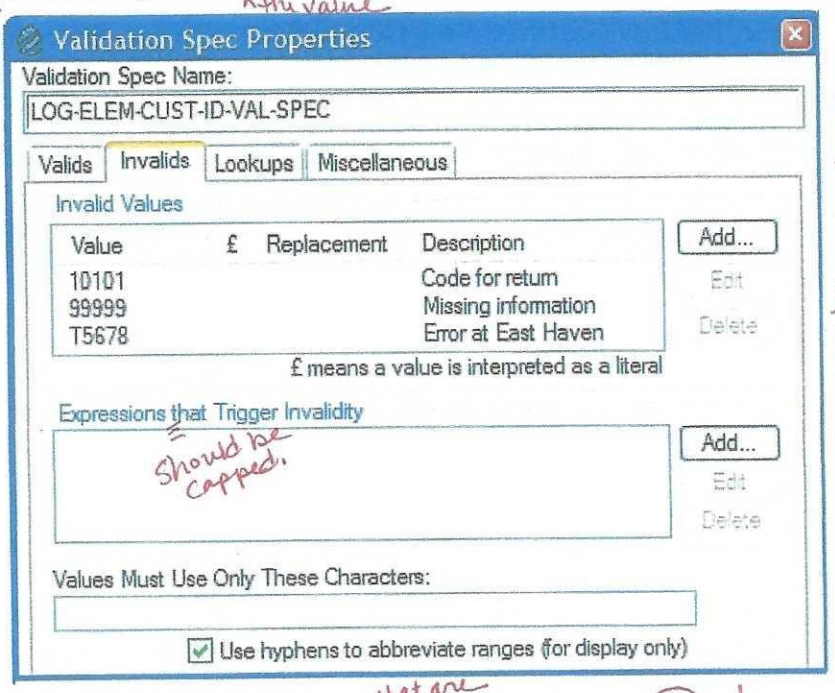
- If a value matches both a valid and an invalid criteria due to contradictory specification in the Validation Spec, the value is considered as to be *invalid*.
- If all four sections of the **Valids** tab of the Validation Spec are blank, and no valid lookup file is specified, *every value is assumed to be valid*, unless it is explicitly *invalid* because of at least one of the following:
  - On the **Invalids** tab of the Validation Spec, the value is in the list of **Invalid Values**, triggers at least one of the **Expressions That Trigger Invalidity**, or contains a character not listed in **Values Must Use Only These Characters**.
  - The value exists in a lookup table of invalid values.
  - The value is illegal for the record format.
- If there is *any* entry in any section of the **Valids** tab, a value is treated as *valid* only if it matches one or more of the entries on the **Valids** tab (and is not explicitly *invalid*).

Invalidity tests

This topic presents a simplified look at the tests that cause a value to be treated as invalid. They are the same tests as those listed in "The Data Profiler tests", but without the details of the validity tests. Use this topic as a quick reference.

These are the Data Profiler invalidity tests listed in the order they are performed:

1. The value is compared against the **Invalid Values** on the **Invalid** tab of the Validation Spec. If it matches any of these, it is treated as *invalid*.



2. If the value contains any characters not listed in **Values Must Only Use These Characters** on the **Invalids** tab, the value is treated as *invalid*.
3. The value is looked up in the lookup file specified under **Invalid Values Lookup** on the **Lookups** tab of the Validation Spec. If the value is found in the invalid lookup file, the value is treated as *invalid*.
4. The value is tested in order against each of the expressions in the **Expressions That Trigger Invalidity** section of the **Invalid** tab. The value is treated as invalid if at least one expression returns a non-zero result. If the expression returns NULL or an error, it is treated as if the value did not trigger the expression.
5. The value is tested by calling the DML function **is_valid**. If is_valid returns **0** (**False**), the value is treated as *invalid*.
6. If any valid value, valid range, valid pattern, valid expression, or valid lookup is defined but the value did not trigger any of those tests, the value is treated as *invalid*.

Another definition of invalid

Here is a definition of *invalid* that differs in presentation (not in content) from those presented in "Invalidity tests" and "The Data Profiler tests". Use whichever definition is the easiest to understand.

The Data Profiler tests values for validity.
The conditions are listed in the same order that the Data Profiler considers them.

~~A value is invalid if at least one of the following are true:~~

- The value is in the list of invalid values.
- Contains any characters not listed in the set of valid characters (but only if a set of valid characters is specified).
- Matches a Key of a lookup table of invalid values.
- Any of the invalidity expressions evaluates to **True** for that value.
- The value is not valid according to the record format (specifically if the **is_valid** DML function returns **False**), and the value is not flagged **ignore in statistics** in the object's Validation Spec.
- The value is not listed as valid ~~(where *valid* means that at least one of the following is True)~~:
  - The value is in the list of valid values.
  - Falls within one of the valid ranges.
  - It's pattern is in the list of valid patterns.
  - Any of the validity expressions evaluate to **True** for that value.
  - The value matches a Key of a lookup table of valid values.
  - There is no explicit definition of *valid* (no valid values, no valid ranges, no valid patterns, no expression to indicate validity, and no lookup table specified for identifying valid values).

■ **NOTE:** ~~The following condition makes a value *neither valid nor invalid*:~~ the record format specifies NULL for a particular value, and the value found during profiling is equal to that value. See "Interpreting a value as NULL".

Why is this section necessary? It's already in "Invalidity tests".

Misleading title. Not another definition, just another way to present the information.

All the information should be easy to understand.

tests for

If a value meets any of the following conditions, the value is invalid.

Should the NULL test be listed? Is the note at the bottom of the page covering NULL?

The value that are

The value in?

u.c. OK?

bold OK?

as

which

bold

OK? Not Ignore this value when computing statistics?

second-level symbol possible?

The value is not

The value's not

does not in?

u.c. OK?

False?

This list confuses me. Why list the valid conditions when the point is to discuss conditions that cause a value to be invalid?

Any value or the Physical Element Value?

If

the testing?

the value is not invalid or valid.

Not in the doc. OK?

About the Data Profiler

The Data Profiler performs tests to determine the invalidity and validity of a Physical Element value. If a value matches one of the tests, the test name appears in the profile results. If a value matches multiple tests, the test name appears in the profile results with one of the descriptions in "Values that match more than one test."

This is my suggestion for a new topic. I'm not sure if it's technically accurate. I took the information from the introduction in "The Data Profiler tests." If the new topic is included, I would delete much of the information in the introduction for "The Data Profiler tests" topic and replace it with a simple sentence that introduces the tests. I would also move the specifics about the information in the profile results to the appropriate topic – "Invalidity tests" – and to "Validity tests" (if written and the profile results information also applies to validity tests).