

# 12

## The Data Profiler tests for valid and invalid values?

to match the order in the doc.  
↓  
invalid and

Sentence style OK for chapter title?

- 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

Not in the doc.

Not in the doc.  
Could be "Both valid and invalid" or

is that a completely different topic?

But, not a head anyway.

If a chapter combines conceptual and reference information; typically, the reference information is placed at the end of the chapter.

Suggest a new organization:

- About the Data Profiler (see suggested topic)
- Important assumptions about validity
- Invalidity tests
- Validity tests (to be written)
- Values that match more than one test
- Data Profiler test descriptions (changed from "The Data Profiler tests")

The tests are about the Physical Element value.

Does the Data Profiler check only for the validity/invalidity of the Physical Element value?

The title should reflect the content.

As written, the title suggests that the Data Profiler checks more than the validity/invalidity of more than one type of value.

Because there 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?

# The Data Profiler tests

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

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".

Latin abbreviations OK?

Tests for invalid values  
Test: NULL

What is the test?  
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".

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

Test: Invalid value

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

Test: Invalid value

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.

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:

**Add Valid Value**

Value: \_\_\_\_\_ Base type: \_\_\_\_\_

☐ Interpret this value as a DML literal

☒ Ignore this value when computing statistics

Description: \_\_\_\_\_

OK Cancel Help

Capping is inconsistent. Do these values need to be capped? Mix of singular and plural OK? Not sure what this is. Add a descriptor?

Statistics?

holding capping OK?

Why is this information here? The test information is reference material, but this information seems to be more procedural (how to).

### Test: is\_valid

There's a test called is-invalid and a function with the same name?

The Physical Element value is tested by calling the the DML function **is\_valid**. If **is\_valid** returns 0 (False), the value is treated as **invalid**. What calls the function? bold OK? function?

■ **NOTE:** If the Physical Element DML definition contains the **is\_valid\_XXX** field validity function, the **is\_valid\_XXX** function is currently ignored. field validity bold OK?

### Test: Valid value

The Physical Element value is compared against 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**. Not sure why this information is relevant. The Physical Element DML function is called. What does the DML definition have to do with the function? Physical Element

### 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**. against or

### Test: Valid pattern

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

Consider: If the Physical Element value matches a valid pattern in the Validation Spec, the Physical Element value is valid.

Contractions aren't being used elsewhere

## Test: Valid expression

Values for expressions that return NULL or an error are ignored?

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.

I'm not sure what the index has to do with the test.

## Test: Valid lookup

Physical Element is compared against

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

matches a value

## Values that match more than one test

Make a head!

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 cited in the Data Profiler tests):

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

Do these show up in the Data Quality Summary and Common Value report?

Inputs, plural.

plural OK?

This is conceptual information. Does it belong with reference material?

The tests produce statistics? On pg. 48, DESCRIPTION is in the Common Value report and REASON is in the Data Quality Summary. Which is correct?

Does DESCRIPTION equal REASON? I don't understand the correlation between the Data Quality Summary and Common Value report.

Bold OK?

## Valid

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

invalid?

## 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**.

verified?

I don't understand what profiling is. OK to use?

There can be multiple Physical Element values in a record.



- 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 more than one test for *validity*. This can happen when multiple validity expressions refer to fields in the 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 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 mean 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 Invalid?

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**.

Should be capped, but may be this way in the UI.

The Data Profiler tests for valid and invalid values

51

Is this really Ignore this value when computing statistics?

*Per the mini doc*

*This information seems out of place.*

*Assuming this is technically correct.*

*Is the description really the first reason?*

*Why bold?*

*Not in the doc. Could it be "Both valid and invalid"?*

*can it be both?*

*Are there the settings?*

*Are there other settings?*

*Wondering if the cross-reference is necessary.*

*bold OK?*

*bold OK?*

*true? So I can't have the reason in the preceding paragraph and one result?*

*Be specific. and?*

*How can a value pass is-valid but fail validity?*

*Is this a must? So the value doesn't match any of the items and passes is-valid and is not flagged. Yes?*

*I'd prefer to spell out the description. What if the preceding paragraph is moved or removed?*

*but as*

*AB INITIO CONFIDENTIAL AND PROPRIETARY — DO NOT COPY*

# Important assumptions about validity

I would avoid including the number of items, to avoid maintenance issues or errors in counting.

Possible to use a second-level symbol?

three human trait - OK?

Conceptual information seems to be general information about validity, but it seems like the Data Profiler checks the

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 **Valid**s 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 **Invalid**s tab of the Validation Spec, the value is in the list of **Invalid Values**, the value 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 **Valid**s tab, a value is treated as *valid* only if it matches one or more of the entries on the **Valid**s tab (and is not explicitly *invalid*).

Not capped in the UI

Considered?

the value

that is

Physical Element values, not all values.

The Physical Element values isn't mentioned in this topic.

the value



There should be a corresponding topic about validity tests.

# Invalidity tests

This seems to be general information about invalidity tests, but the tests are for the physical Element value - seeing that they are the same tests listed in "The Data Profiler 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.

For more information about the tests, see "The Data Profiler tests."

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.

Should the null test be listed? It's considered a test for invalidity.

the value

Value	Replacement	Description
10101		Code for return
99999		Missing information
T5678		Error at East Haven

£ means a value is interpreted as a literal

Expressions that Trigger Invalidity

Values Must Use Only These Characters:

☒ Use hyphens to abbreviate ranges (for display only)

Not sure why this image is placed here. It would make more sense to me to have the image before the introduction that tells the user it's an example of what the user will see in the UI.

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.

I don't understand "treated as if the value did not trigger the expression". It sounds like something is being ignored, but I don't know what.

which tests? for valid value etc?

Doesn't match the UI, but the word should be capped.

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

## Another definition of invalid

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

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.
- <sup>The value</sup> Contains any characters <sup>that are</sup> not listed in the set of valid characters (but only if a set of valid characters is specified).
- <sup>The value</sup> Matches a Key <sup>in?</sup> 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 <sup>not</sup> in the list of valid values.
  - <sup>The value is not</sup> Falls within one of the valid ranges.
  - <sup>The value's</sup> It's pattern is in the list of valid patterns.
  - Any of the validity expressions evaluate to **True** for that value.
  - The value <sup>does not</sup> matches a <sup>in?</sup> 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".

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

Second-level symbol possible?

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 on the Physical Element value?

OK? Not  
Ignore this value when computing statistics?

If  
the value is not invalid or valid.

Not in the dec. 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).