

教学质量监控数据治理项目
数据质量稽核报告

报告日期：2025.8.18

目录

一、数据质量管理范围、目标与流程	1
(一) 稽核范围	1
(二) 稽核目标	1
(三) 稽核流程	1
1、数据质量稽核规则设计	1
2、数据标准配置	2
3、数据质量稽核作业配置	2
4、数据清洗规则制定	2
(四) 数据质量评分等级	3
二、数据质量评估结果分析	3
(一) 数据质量问题分析	3
1、系统级数据问题分析	4
2、表级数据问题分析	4
(二) 数据质量问题分析与处理	5
1、无问题的数据	5
2、问题已解决的数据	6
3、待技术解决的问题数据	6
4、待业务方确认的问题	9
附件 1 数据质量稽核范围	12
附件 2 数据质量稽核规则	12
附件 3 无问题的数据	12
附件 4 八年内无问题的数据	12
附件 5 问题已解决的数据	13
附件 6 待技术解决的问题数据	13
附件 7 待业务方确认的问题	13

一、数据质量管理范围、目标与流程

（一）稽核范围

本次数据稽核范围，为“教学质量监控数据治理项目”中，指标计算用到的 20 张数据表和 59 个非主键的计算字段，涉及国家开放大学体系内的招生系统、学习网、教务系统和考务系统，具体表及字段信息见附件 1。

（二）稽核目标

围绕质量因子指标相关数据表，从完整性、准确性、一致性等维度全面评估数据质量，精准识别问题数据并分类解析成因，据此制定针对性的数据清洗策略，形成“评估-诊断-治理”的质量提升闭环。

（三）稽核流程

1、数据质量稽核规则设计

数据质量稽核规则体系设计以保障指标计算准确性 20 张关键数据表及 59 个字段的多维度拆解，构建包含 87 条量化规则的稽核体系，通过明确指标计算的字段映射关系与校验逻辑，实现数据质量问题的精准定位与高效治理。

基于数据质量“六性”（完整性、准确性、一致性、及时性、唯一性、有效性）的要求，结合字段的业务定义、数据流向及应用场景特性，构建覆盖业务规则校验、技术逻辑验证的多层级数据质量稽核规则体系。稽核规则见附件 2。

2、数据标准配置

基于数据质量稽核规则体系，在数据治理平台的数据标准模块中，为被稽核字段构建标准化稽核基准。针对格式校验与值域校验场景，在数据标准模块中配置字段标准格式及值域正则表达式；针对参照校验场景，在码表管理模块维护标准化码表体系，并在数据标准模块建立字段与码表的引用映射关系。本次治理共完成 29 条正则表达式（格式校验）规则构建、20 张标准化码表（参照校验）维护，形成可复用的数据质量稽核标准库。

3、数据质量稽核作业配置

在数据治理平台的质量作业模块中，以数据表为单元构建字段级稽核任务体系，通过配置动态调度策略实现稽核作业的自动化执行，最终生成表级与字段级的多维度质量评分，并将问题数据结构化沉淀至数仓问题库。本次治理已完成 20 个质量稽核作业的体系化配置，覆盖 59 个字段的 87 条稽核规则，形成"任务配置-自动执行-质量评估-问题沉淀"的管理机制。后续若该 20 张数据表产生新增字段稽核需求，可直接在现有作业体系中动态追加规则，无需重复创建作业实例，显著提升数据治理效率。

4、数据清洗规则制定

组建由业务部门、数字化部、数据治理团队与系统研发团队构成的跨领域数据质量攻坚专班，通过逐字段解析

问题数据特征，针对性制定多维度数据清洗策略。经攻坚分析，问题数据可划分为三类典型场景：1）可直接清洗的：通过数据清洗工具或脚本规则能够直接修正的字段异常；2）需要业务系统改造的：因业务系统功能缺陷或数据采集逻辑漏洞导致的源头性数据问题；3）需要业务部门协助确认的：涉及业务定义模糊或历史数据遗留等问题，需联合业务部门进行规则澄清或人工核验的数据问题。

（四）数据质量评分等级

构建数据质量评分模型，采用缺陷率倒数法计算字段合格率，作为评分依据。

$$\text{字段合格率} = (\text{总记录数} - \text{问题记录数}) / \text{总记录数} \\ * 100\%$$

例：若问题记录数占比 24.39%，则合格率为 75.61%，直接映射为百分制评分 75.61 分。

表级评分采用各字段评分的算术平均值作为计算基准。

基于百分制评分，构建数据表三级质量评估体系：

- （1）优秀：100 分
- （2）良好：95-99 分
- （3）不及格：95 分以下

二、数据质量评估结果分析

（一）数据质量问题分析

1、系统级数据问题分析

全量数据和八年内数据稽核结果均显示，参与数据质量稽核的四个业务系统中，招生系统评分最高，其余依次是学习网、教务系统和考务系统。

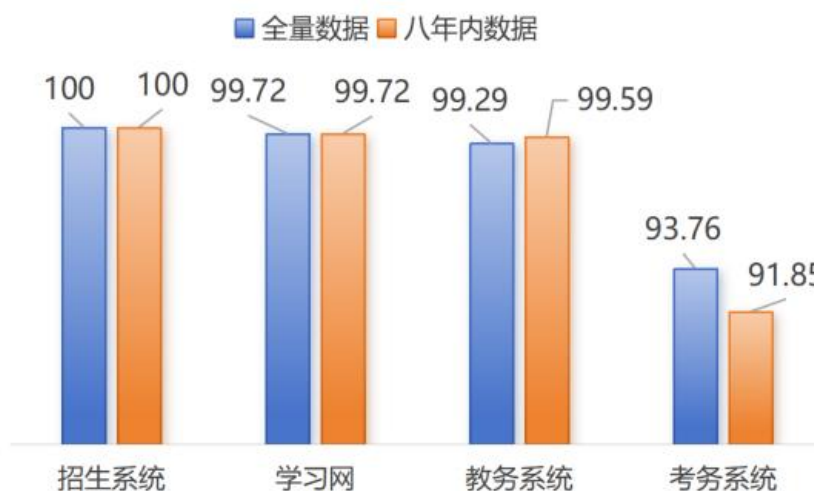


图 1 各业务系统数据质量评分（全量数据&八年内数据）

注：评分仅代表参与质量因子计算的部分表和部分字段情况，不代表业务系统的整体数据质量水平。

2、表级数据问题分析

本次数据质量稽核覆盖 20 张业务数据表，全量数据和八年内数据稽核结果均显示：优秀等级表 13 张（占比 72.22%），良好等级表 7 张（占比 38.89%），不及格表 1 张（占比 5.56%）。其中，优秀等级表的高占比体现数据质量具备较好基础，但良好与不及格表合计占比 44.44%，反映数据质量仍存在较大的提升空间。因此，需要针对此类表开展字段级质量溯源分析，聚焦完整性缺失、格式不

正确、值域异常等核心缺陷，构建分层提升策略，通过规则迭代、系统改造及业务协同等多元手段，持续夯实数据质量基底，为教学质量监控等业务场景提供更具可信度的数据支撑。

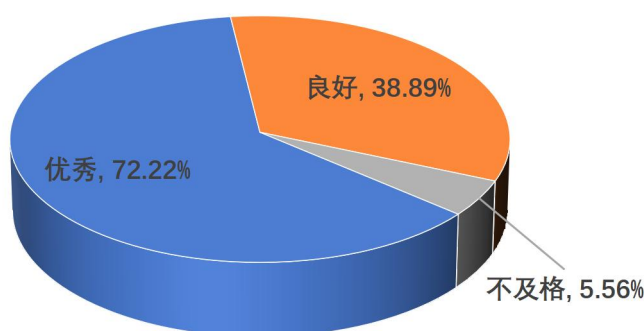


图2 各评分等级表数量占比(全量数据&八年内数据)

(二) 数据质量问题分析与处理

本次对20张业务表的59个字段执行87条稽核规则，从全量数据来看，51条稽核规则验证通过，占比58.62%；36条规则触发异常，占比41.38%。从八年内数据来看，59条稽核规则验证通过，占比67.82%；26条规则触发异常，占比32.18%。涉及值域异常、格式错误、数据缺失、枚举值异常、数据不一致五类典型错误，根据问题确认和解决情况，共分为以下四类：无问题、已解决、待技术解决、待业务方确认。

1、无问题的数据

稽核结果显示，51条稽核规则全量数据校验通过，涉及

18 张表的 41 个字段。8 条规则仅八年内数据校验通过，涉及 4 张表的 8 个字段。全量校验通过的字段及规则见附件 3。八年内无问题的字段及规则见附件 4。

2、问题已解决的数据

19 条规则稽核出的数据问题已解决，涉及 11 张表的 15 个字段。其中，各科室已解决条数和问题总条数、遗留问题情况如下：

1) 过程科 (7/10)：遗留问题 2 条待技术解决，1 条待业务方确认；

2) 学籍科 (6/11)：遗留问题 5 条待业务方确认；

3) 学位办 (1/1)：无遗留问题；

4) 考务科 (1/11)：遗留问题均为待技术解决；

5) 招办 (4/5)：遗留问题 1 条待业务方确认。

6) 学科科 (2/2)：无遗留问题。

问题已解决的字段及规则、问题描述见附件 5。

3、待技术解决的问题数据

11 条规则稽核出的数据问题待技术解决，涉及 3 张表的 6 个字段。待技术解决的字段及规则、问题描述见附件 6。以下是待技术解决的问题描述：

(1) 学习网

1) 学习网上传参考信息表“上级类型”字段，存在 1136 条空值，需要学习网增加程序，及时自动删除同步任

务失败重试操作导致的问题数据记录。**最新进展（2025.8.1）**：学习网采用弹性的部署方式。访问量大会自动增加容器数量，访问量小会自动收缩容器数量。少数异步任务在执行期间，碰到容器收缩导致异常退出。正在进行异常处理的优化工作。截止到当前时间8月1日，因并行处理学习网暑假直播以及样板课功能升级改造，暂未完成处理优化。正在进行中，预计9月底完成，异步任务比较多，每处理完一个就上线一个。

2) 学习网考试测验成绩信息表“原始成绩”字段，存在5814条大于100的记录，经排查为刷课导致，举例：客观题10道，学生提交了11道，算分时会多。不影响评阅率。已增加大于100分数重新核验机制，后续需要增加提交校验。**最新进展（2025.8.1）**：如果不临时插入优先级更高的工作，预计9月底完成。

（2）考务系统

1) 考务系统考试成绩明细表“形考成绩”字段，在“形考比例”字段大于0的情况下，该字段有约7%的记录为空。初步分析这部分形考记录表示缺考，待技术进一步确认，预计10月初开始排查原因和解决。

2) 考务系统考试成绩明细表“形考成绩”字段，“形考比例”字段为空或0，该字段存在有效值。经分析，原因系成绩从历史系统、学习网同步过来导致，统计时可以

不做计算。待技术在数据同步的时候加校验机制，预计10月初开始解决。

3) 考务系统考试成绩明细表“终考成绩”字段，存在47000+小于0记录，目前无法找到小于0的原因，需要后续再出现小于0的情况时再做分析。

4) 考务系统考试成绩明细表“终考成绩”字段，“形考比例”字段为空或小于100，该字段存在约8%的空值。后续需确认除了标记缺考的，数据为空值的原因，考虑两种情况：来参加考试刷脸+登录+交白卷，来参加考试刷脸没登录，预计10月初开始排查原因。

5) 考务系统考试成绩明细表“终考成绩”字段，“形考比例”字段为100，该字段有约1%的记录不为空。需要加校验机制，终考成绩不为空的情况下限制录入或同步终考成绩，预计10月初开始解决。

6) 考务系统考试成绩明细表“终考录入码表”字段，存在约60%的记录不符合码值要求。需要技术排查原因，预计10月初开始排查。

7) 考务系统考试成绩明细表“终考录入码表”字段，“形考比例”字段为空或小于100，该字段存在约8%的空值记录。待技术排查原因，预计10月初开始排查。

8) 考务系统考试成绩明细表“合成成绩”字段，存在47187条小于0的记录。待技术排查原因，预计10月

初开始排查。

9) 考务系统考试成绩明细表“合成成绩”字段，存在约 10%的空值记录。待技术排查原因，预计 10 月初开始排查。

4、待业务方确认的问题

6 条规则稽核出的数据问题待业务方确认，涉及 3 张表的 4 个字段。其中，过程科 1 条，学籍科 5 条，招办 1 条（需要和学籍科共同确认）。待业务方确认的字段及规则、问题描述见附件 7。

以下是待确认问题的核心内容简述：

(1) 过程科

学习网课程信息表“学期课程分部代码”字段，**数据格式问题**：存在缺少位数的字符串，以及英文和特殊字符的组合，数字和特殊字符的组合。经分析，不符合标准的内容均为课程开放的课程代码。**当前确认结论是**：可以在表里增加字段区分课程开放和学历教育，过程科需要和系统研发进一步研究决定是否区分课程开放与学历教育，评估可行性与性价比。

(2) 学籍科

1) 教务系统学籍信息表“毕业年度学期”，**数据格式问题**：22 秋及之前数据中，2706536 条格式为：年份+春季、年份+秋季；752185 条格式为：年份+01、年份+07。

当前确认结论是：已提供明细数据，供学籍科进一步分析后确定清洗方案。

2) **教务系统学籍信息表“毕业年度学期”，数据一致性问题：**该字段有效的情况下，学籍状态字段内容不是毕业。**当前确认结论是：**已提供问题数据明细，包括对应的分部，供学籍科进一步确认清洗方案。

3) **教务系统学籍信息表“毕业年度学期”，数据缺失问题：**学籍状态为毕业的情况下，有一条空值，对应的学生信息如下：学号 0712001200160，分部：120，学习中心：1204444。**当前确认结论是：**预计是 CPS 数据迁移时丢失，待学籍科进一步确认清洗方案。

4) **教务系统学籍异动信息表“申请学年学期”数据格式问题：**61 条没有显示年份信息。**当前确认结论是：**已提供问题数据异动前、异动后所属分部、姓名、学号的信息，供学籍科进一步确认原因和清洗方案。

(3) 学籍科和招办协商处理

教务系统学籍信息表“学号”，数据格式问题：三条不符合格式要求（96127303、9944001413442、9951250001）。**当前确认结论是：**学籍科和招办协商处理。

附件 1 数据质量稽核范围



附件1
数据质量稽核范围

附件 2 数据质量稽核规则



附件2
数据质量稽核规则

附件 3 无问题的数据



附件3
无问题的数据.xlsx

附件 4 八年内无问题的数据



附件4
八年内无问题数据

附件 5 问题已解决的数据



附件5
问题已解决的数据

附件 6 待技术解决的问题数据



附件6
待技术解决的问题

附件 7 待业务方确认的问题



附件7
待业务方确认问题