# DESARROLLO Y VALIDACIÓN PSICOMÉTRICA DE UN PROTOTIPO DE PRUEBA OBJETIVA PARA EVALUAR LA ALFABETIZACIÓN VISUAL EN ESTUDIANTES DE ÚLTIMO SEMESTRE DEL PROGRAMA DE DISEÑO DIGITAL Y MULTIMEDIA DE LA UNIVERSIDAD COLEGIO MAYOR DE CUNDINAMARCA

Héctor Castelblanco Mendieta

Maestría en educación.

Facultad de Ciencias de la Educación

Universidad la Gran Colombia

Bogotá D.C.

2026

# Desarrollo y Validación Psicométrica de un Prototipo de Prueba Objetiva para Evaluar la Alfabetización Visual en Estudiantes de Último Semestre del Programa de Diseño Digital y Multimedia de la Universidad Colegio Mayor de Cundinamarca

Héctor Castelblanco Mendieta

Trabajo de grado presentado como requisito para optar al título de

Magíster en Educación

Yudi Andrea Ortiz Rocha

Cargo Directora de tesis

Doctora en Ciencias en la Especialidad de Matemática Educativa

Universidad La Gran Colombia

Facultad de Ciencias de la Educación

Maestría en Educación

Bogotá D.C., 2026

Bogotá D.C., 20 de abril de 2026

Señores
Departamento de Biblioteca
Universidad La Gran Colombia
Ciudad

Yo, Héctor Castelblanco Mendieta, identificado(a) con cédula de ciudadanía número 79318786, autor del trabajo de grado titulado Desarrollo y validación psicométrica de un prototipo de prueba objetiva para evaluar la alfabetización visual en estudiantes de último semestre del Programa de Diseño Digital y Multimedia de la Universidad Colegio Mayor de Cundinamarca, presentado como requisito para optar al título de Magíster en Educación de la Universidad La Gran Colombia, autorizo al Departamento de Biblioteca de la Universidad La Gran Colombia para que, con fines académicos, realice la consulta parcial o total de este trabajo de grado, así como su consulta o reproducción parcial o total, o la publicación electrónica del texto completo en el repositorio institucional y su registro en el catálogo KOHA del Departamento de Biblioteca.

El autor declara que la obra es original y que no vulnera ningún derecho de terceros.

Atentamente,

Héctor Castelblanco Mendieta

Programa Maestría en Educación

**Resumen**

Esta investigación de tipo instrumental desarrolló y validó un prototipo de prueba objetiva para evaluar la alfabetización visual en estudiantes de último semestre de programas de diseño. El estudio responde a la carencia concreta de instrumentos estandarizados capaces de medir con rigor psicométrico estas competencias visuales. El constructo se fundamentó en los estándares de la *Association of College and Research Libraries* (ACRL, 2011) y se estructuró mediante la metodología del Diseño Centrado en Evidencias, produciendo un instrumento de 34 ítems de selección múltiple. La validación adoptó una estrategia de triangulación convergente que ejecutó simultáneamente la evaluación de contenido por juicio de seis expertos y una prueba piloto con 54 estudiantes de último semestre del programa de Diseño Digital y Multimedia de la Universidad estatal colombiana Colegio Mayor de Cundinamarca. Los resultados confirman la viabilidad técnica del instrumento. La validez de contenido obtuvo coeficientes V de Aiken superiores a 0.90 en relevancia y coherencia. El análisis bajo la Teoría Clásica de los Tests confirmó la unidimensionalidad esencial del constructo y reportó una consistencia interna aceptable (KR-20 = 0.758). La calibración bajo el Modelo de Rasch mostró que los 32 ítems depurados se ajustan al modelo probabilístico sin ruido estadístico significativo. Se detectó un efecto techo que refleja la consolidación curricular de las competencias técnicas y éticas en la cohorte evaluada, y señala la lectura visual crítica como la dimensión de mayor variabilidad en el perfil de egreso.

*Palabras clave:* Alfabetización visual · Validación de instrumentos · Diseño Centrado en Evidencias · Modelo de Rasch · Teoría de Respuesta al Ítem · Formación en diseño · Educación superior en Colombia.

**Abstract**

This instrumental research developed and validated a prototype objective test to assess visual literacy in final-semester students of design programs. The study addresses the concrete lack of standardized instruments capable of measuring these visual competencies with psychometric rigor. The construct was grounded in the standards of the Association of College and Research Libraries (ACRL, 2011) and structured following the Evidence-Centered Design methodology, resulting in a 34-item multiple-choice instrument. Validation adopted a convergent triangulation strategy, simultaneously conducting content validity by six expert judges and a pilot test with 54 final-semester students from the Digital and Multimedia Design program at the Colombian state university Colegio Mayor de Cundinamarca. The results confirm the technical feasibility of the instrument. Content validity yielded Aiken's V coefficients above 0.90 for relevance and coherence. Analysis under Classical Test Theory confirmed the essential unidimensionality of the construct and reported acceptable internal consistency (KR-20 = 0.758). Calibration under the Rasch model showed that the 32 refined items fit the probabilistic model without significant statistical noise. A ceiling effect was detected, reflecting the curricular consolidation of technical and ethical competencies in the evaluated cohort, and identifying critical visual reading as the dimension with the greatest variability in the graduate profile.

*Keywords*: Visual literacy · Instrument validation · Evidence-Centered Design · Rasch model · Item Response Theory · Design education · Higher education in Colombia.

**TABLA DE CONTENIDO**

# ÍNDICE DE TABLAS

## ÍNDICE DE FIGURAS

# INTRODUCCIÓN

La imagen ocupa un lugar central en los procesos de comunicación contemporáneos. Avgerinou y Pettersson (2011) documentan que lo visual ya no funciona únicamente como registro de la realidad, sino como medio activo para construirla y transformarla. En los programas de formación en diseño, esta realidad tiene implicaciones directas sobre el perfil de competencias que los egresados deben demostrar: interpretar, producir y evaluar críticamente imágenes no es una habilidad complementaria, sino el eje del ejercicio profesional. Lo que la educación superior no ha resuelto es cómo verificar ese dominio antes de la graduación. No existen instrumentos estandarizados con fundamento psicométrico que permitan medir si los estudiantes de diseño alcanzan los niveles de alfabetización visual que su perfil de egreso declara. Esta investigación desarrolla y valida un primer prototipo de ese instrumento.

A pesar del consenso global sobre la importancia de la alfabetización visual, la cual ha sido fundamentada entre otras, por organizaciones como la *Association of College and Research Libraries* (ACRL), en la evaluación de la educación superior subsiste una brecha metodológica profunda: aunque se enseña a leer, escribir y pensar con imágenes, todavía no se dispone de herramientas estandarizadas para medir objetivamente los niveles de competencia visual de los estudiantes. Aunque la evaluación en diseño se ha apoyado históricamente en portafolios y juicio de docentes expertos, estas herramientas resultan insuficientes para certificar con evidencia empírica objetiva el dominio de competencias visuales de sus egresados a escala institucional (ACRL, 2011; Kędra, 2018).

La irrupción reciente de la inteligencia artificial generativa en los procesos de producción y consumo visual intensifica esta carencia. Brumberger (2025) argumenta que la IA generativa redefine las competencias requeridas para leer y producir imágenes, ampliando el repertorio de

conocimientos necesarios para detectar manipulaciones, evaluar sesgos algorítmicos y discernir entre contenido auténtico y sintético.

Esta investigación nace ante la necesidad de desarrollar un instrumento de medición psicométrica estandarizada de la alfabetización visual en la educación superior en Diseño. El presente documento presenta el desarrollo metodológico, técnico y analítico que se realizó para validar un prototipo de prueba objetiva enfocada en evaluar las competencias en alfabetización visual en estudiantes de último semestre del programa de Diseño Digital y Multimedia, de la institución estatal de educación superior colombiana, Universidad Colegio Mayor de Cundinamarca.

El documento se estructura en cinco capítulos. El primero sitúa la problemática en sus dos dimensiones: la brecha global en la evaluación estandarizada de la alfabetización visual y la carencia específica del sistema colombiano de evaluación de la educación superior, cuyas pruebas de Estado no incluyen indicadores para medir las competencias de interpretación crítica, capacidad técnica y uso ético de contenidos visuales en los programas de diseño. (ICFES, 2018, 2019). A partir de ese diagnóstico, el capítulo formula la pregunta de investigación y los objetivos del estudio. El segundo construye el andamiaje conceptual del instrumento: traza la evolución del concepto de alfabetización visual desde sus formulaciones iniciales (Debes, 1969) hasta el modelo de tres dimensiones interdependientes adoptado como referencia estructural —lectura, escritura y ética visual— (Kędra, 2018). Presenta los siete estándares de la ACRL (2011) como operacionalización del constructo y explica cómo el Diseño Centrado en Evidencias (Mislevy et al., 2003) permite traducir esos estándares en indicadores observables y medibles. El capítulo cierra con la descripción de los marcos psicométricos del estudio —la Teoría Clásica de los Tests y el Modelo de Rasch— que articulan la estrategia de validación. El tercero describe las decisiones

metodológicas que convirtieron ese marco conceptual en un instrumento concreto: el enfoque cuantitativo de tipo instrumental, el diseño de validación convergente por triangulación metodológica (Kane, 2013), el proceso iterativo de construcción en cuatro versiones, el uso de herramientas de inteligencia artificial generativa para la producción de estímulos visuales controlados y los procedimientos del panel de jueces y del análisis psicométrico. El cuarto expone y discute los resultados de la validación convergente. Examina la validez de contenido establecida por el panel de expertos, la estructura interna del instrumento derivada de los datos del piloto —incluyendo los análisis de dificultad, discriminación y unidimensionalidad—, la calibración exploratoria bajo el Modelo de Rasch y el contraste entre la dificultad teórica prevista y la empíricamente observada. La sección de triangulación articula ambas fuentes de evidencia e identifica sus convergencias, sus divergencias y las implicaciones de cada una para el desarrollo del instrumento. El quinto sintetiza las conclusiones, examina los hallazgos de mayor alcance transversal y traza la ruta de desarrollo que el instrumento requiere para consolidarse como herramienta de uso institucional. Precisa el alcance y las condiciones concretas del estudio y propone lineamientos técnicos para las fases sucesivas de depuración, calibración definitiva e implementación a escala.

El recorrido por estos cinco capítulos traza el camino desde la carencia —la ausencia de herramientas estandarizadas para medir lo que el diseño enseña— hasta una respuesta técnicamente fundamentada. Esta investigación ofrece a las instituciones de educación superior una ruta inicial para verificar, con criterio estandarizado, que están formando diseñadores capaces de leer, crear y usar imágenes con responsabilidad técnica y ética.

# CAPÍTULO 1. PROBLEMÁTICA

En la educación superior contemporánea, la alfabetización visual se ha consolidado como una competencia esencial en los programas de formación en diseño: en un entorno donde la cultura visual y la digitalización de la información son predominantes, la capacidad de interpretar, analizar y producir contenido visual define las condiciones mínimas del ejercicio profesional competente (Hattwig et al., 2013).

En este contexto, y de acuerdo con los lineamientos de la *Association of College and Research Libraries* (ACRL, 2011), la alfabetización visual se define como el conjunto de competencias que permiten a los individuos interpretar, analizar, evaluar, crear y utilizar imágenes de manera crítica, ética y efectiva en distintos contextos académicos, sociales y culturales. Este dominio abarca tanto la decodificación visual —la comprensión de significados, estructuras y contextos— como la codificación visual —la capacidad para diseñar y comunicar mediante recursos gráficos— y el uso ético de materiales visuales (Kędra, 2018; Avgerinou y Pettersson, 2011). Las directrices y estándares de competencias que tipifican y describen estas habilidades fueron establecidas por la *Association of College and Research Libraries* (ACRL) en 2011 y constituyen la base operativa de esta investigación, pues definen el constructo mediante indicadores de desempeño y resultados de aprendizaje estrictamente observables, condición indispensable para el diseño de ítems cerrados y su calibración bajo la Teoría de Respuesta al Ítem (TRI). La justificación detallada de esta elección frente al marco conceptual más flexible, actualizado por la ACRL en 2022 se desarrolla en el Capítulo 2.

Aunque la literatura académica ha avanzado en describir los aportes de la alfabetización visual en los aspectos cognitivos del procesamiento de información, los factores estéticos relacionados con la apreciación y producción formal y los elementos críticos para el análisis ético

e ideológico, persiste una brecha evidente entre los avances teóricos y la práctica evaluativa. Autores como Avgerinou y Pettersson (2011) señalan que la alfabetización visual no se ha reconocido suficientemente como un componente esencial en muchos currículos, lo que ha impedido el desarrollo de políticas claras para su medición objetiva.

Esta insuficiencia de herramientas de evaluación estandarizados representa un problema crítico. La multiplicidad de enfoques teóricos y la falta de consenso sobre los componentes específicos del constructo han dificultado la creación de marcos coherentes para su medición. Como indican Li y Potter (2023), muchas instituciones educativas carecen de metodologías claras para integrar estas competencias, y la rápida evolución tecnológica deja obsoletas las herramientas de enseñanza tradicionales, lo que dificulta la estandarización de enfoques educativos.

A nivel global, aunque se han desarrollado instrumentos como el Visual Literacy Assessment Test (VLAT), la investigación empírica sobre medición de competencias visuales contrasta con la abundancia del acervo teórico. La mayoría de los instrumentos existentes se concentra en evaluar la decodificación pasiva de imágenes —cómo el observador lee, procesa y organiza información visual— dejando fuera las dimensiones que más importan en la formación de un diseñador: la producción visual, el juicio crítico-ético y el análisis de la intencionalidad y los efectos comunicativos de lo visual (Natharius, 2004; Li y Potter, 2023; Pandey y Ottley, 2023; Kędra, 2018; Thompson, 2021). Esta brecha entre la teoría y la medición impide diagnosticar con precisión el nivel de competencia de los estudiantes y orientar con evidencia las decisiones curriculares.

En el contexto latinoamericano y colombiano, la evaluación de competencias visuales ha dependido de enfoques cualitativos como la revisión de portafolios y las rúbricas de aula, estrategias que, aunque valiosas para el seguimiento formativo, no permiten contar con

diagnósticos comparables entre cohortes ni certificación objetiva a escala institucional (Morales, 2022).

A nivel nacional, el ICFES ha consolidado las pruebas de estado Saber Pro como el principal instrumento de evaluación estandarizada de la educación superior, construidas sobre el Diseño Centrado en Evidencias y la Teoría de Respuesta al Ítem (ICFES, 2018). Estas pruebas cubren competencias genéricas como Lectura Crítica y Razonamiento Cuantitativo, y competencias específicas por área de conocimiento. El módulo dirigido a los programas de diseño, denominado Módulo de Generación de Artefactos, evalúa principalmente competencias proyectuales y metodológicas (ICFES, 2019), sin incluir indicadores para medir la interpretación crítica de imágenes, la producción visual estratégica o el uso ético de contenidos visuales.

Estos vacíos del sistema nacional tienen consecuencias concretas: los programas de diseño no disponen de datos estandarizados que les permitan saber si sus egresados dominan competencias tan centrales para su ejercicio profesional como la búsqueda estratégica de imágenes, la interpretación crítica o el uso ético de la información visual, ni identificar en cuáles de ellas el currículo está fallando.

El programa de Diseño Digital y Multimedia de la Universidad Colegio Mayor de Cundinamarca no escapa a esta situación. Su Proyecto Educativo del Programa (PEP) declara que el egresado "aplica el lenguaje específico de la expresión gráfica, representación y visualización de proyectos" y que es “capaz de reflexionar con sentido crítico y conciencia histórica" (Universidad Colegio Mayor de Cundinamarca, 2024, p. 13), competencias compatibles con las dimensiones de escritura visual y pensamiento crítico del modelo de Kędra (2018). Sin embargo, el PEP no formaliza la alfabetización visual como categoría evaluativa ni la operacionaliza mediante indicadores de desempeño verificables. El sistema de evaluación del programa —

articulado mediante rúbricas de proyecto, portafolios, bitácoras de clase y jurados internos y externos— es valioso para el seguimiento formativo, pero descansa enteramente en el juicio de docentes y evaluadores expertos, sin ofrecer diagnósticos comparables entre cohortes ni mediciones escalables a nivel institucional. El programa enseña a leer, producir y usar imágenes, pero carece de un instrumento validado que permita verificar, antes de la graduación, en qué medida sus estudiantes dominan efectivamente esas competencias.

Ante esta carencia, la presente investigación adopta los modelos psicométricos de la Teoría de Respuesta al Ítem (TRI) como alternativa metodológica. A diferencia de los enfoques clásicos, la TRI produce estimaciones de dificultad y habilidad independientes de la muestra y ofrece información diagnóstica ítem a ítem (Börner et al., 2019). Su desarrollo metodológico se detalla en el Capítulo 3.

Sin datos estandarizados sobre los niveles de dominio de las competencias visuales de sus egresados, el programa tampoco puede identificar qué dimensiones del currículo requieren ajuste. De esa carencia surge la siguiente pregunta de investigación:

¿Qué evidencias de validez de contenido y qué propiedades psicométricas presenta un prototipo de prueba objetiva, fundamentado en los estándares de alfabetización visual de la ACRL y estructurado mediante el Diseño Centrado en Evidencias, para evaluar la alfabetización visual en estudiantes de último semestre del programa de Diseño Digital y Multimedia de la Universidad Colegio Mayor de Cundinamarca?

## OBJETIVOS

### OBJETIVO GENERAL

Desarrollar y validar psicométricamente un prototipo de prueba objetiva para evaluar la alfabetización visual en estudiantes universitarios de último semestre, verificando su validez de

contenido, sus propiedades psicométricas bajo la Teoría Clásica de los Tests y el cumplimiento de los supuestos de unidimensionalidad requeridos para su futura calibración paramétrica[1] bajo la Teoría de Respuesta al Ítem.

**OBJETIVOS ESPECÍFICOS**

1- Identificar las dimensiones y competencias clave de la alfabetización visual relevantes para el contexto de formación en diseño digital, a partir de los estándares de la ACRL (2011).

2- Diseñar una prueba objetiva que contemple ítems alineados con las dimensiones identificadas, equilibrando los aspectos de decodificación, codificación y uso ético de imágenes.

3- Verificar la validez de contenido del instrumento mediante la revisión de referentes teóricos, el análisis de instrumentos previos y la evaluación de los ítems por parte de docentes expertos del área de diseño.

4- Evaluar las propiedades psicométricas del instrumento a partir de los datos empíricos recolectados en la prueba piloto, analizando la confiabilidad, dificultad y discriminación bajo la Teoría Clásica de los Test verificando las condiciones requeridas para la Teoría de Respuesta al Ítem.

5- Estructurar una versión depurada del instrumento a partir de la triangulación de evidencias (juicio de expertos y desempeño empírico), proponiendo los lineamientos técnicos necesarios para su futura calibración paramétrica a gran escala.

[1] Calibración paramétrica: procedimiento estadístico mediante el cual se estiman los parámetros de dificultad y discriminación de cada ítem dentro de un modelo probabilístico, de modo que sus propiedades queden expresadas en una escala común e independiente de la muestra específica utilizada. Adaptado de Bond y Fox, 2015.

# CAPÍTULO 2. MARCO CONCEPTUAL

## 2.1. Origen y desarrollo de la Alfabetización visual

Cuando John Debes acuñó el concepto en la Primera Conferencia Nacional sobre Alfabetización Visual celebrada en Rochester en 1969, lo hacía pensando en la capacidad humana básica para pensar, aprender y comunicarse a través de estímulos visuales (Debes, 1969). Esta definición era esencialmente perceptiva. Lo que ese momento no podía anticipar era la velocidad con la que el concepto de alfabetización visual se tornaría en una noción más compleja y fundamental para la sociedad contemporánea.

En las cinco décadas siguientes, el avance de las tecnologías digitales y la consolidación de una cultura visual global transformaron el constructo de la alfabetización visual enriqueciéndolo y ampliándolo con nuevas dimensiones. Avgerinou y Pettersson (2011) lo sintetizan con precisión: lo visual ya no solo documenta o representa la realidad, sino que tiene el poder de construirla y transformarla. Esta constatación desplazó el foco desde la percepción hacia la competencia activa. Müller (2008) propuso organizar esa competencia en cuatro dimensiones —producción, percepción, interpretación y recepción— marcando el paso de una noción pasiva a una concepción integral que exige al profesional visual una postura analítica, propositiva y éticamente responsable frente a la imagen.

Desde esa base, Kędra (2018) formuló el modelo que esta investigación adopta como referencia estructural. La competencia visual del profesional descansa, en su propuesta, sobre tres dimensiones interdependientes. La lectura visual es la capacidad hermenéutica[2] de interpretar significados latentes, analizar estructuras compositivas y comprender los contextos históricos, culturales y políticos en que las imágenes se producen. La escritura visual va más allá de la técnica:

[2] Hermenéutica visual: práctica interpretativa orientada a descifrar los significados, las intenciones y los contextos culturales que subyacen a una imagen, más allá de su lectura literal o descriptiva. Adaptado de Kędra, 2018.

abarca la habilidad de generar comunicaciones visualmente coherentes y éticamente fundamentadas, tomando decisiones deliberadas sobre qué recursos gráficos usar y qué impacto producir en una audiencia concreta. El pensamiento visual, por su parte, es el uso estratégico de representaciones gráficas para estructurar razonamientos, resolver problemas abstractos y construir conocimiento nuevo.

La mayoría de los instrumentos de evaluación existentes no miden las tres dimensiones de manera equivalente. Li y Potter (2023) advierten que la medición de la alfabetización visual se ha enfocado principalmente en la lectura visual, dejando fuera las dimensiones más críticas para el ejercicio profesional de un diseñador: la escritura visual y el juicio ético. Para los programas que forman profesionales cuyo trabajo consiste en tomar decisiones deliberadas sobre la forma, el significado y el uso legal de las imágenes, un instrumento que no evalúe esas dimensiones es, en el mejor de los casos, insuficiente.

Para delimitar el dominio del constructo y corregir esas carencias, esta investigación adopta los Estándares de Competencia en Alfabetización Visual para la Educación Superior formulados por la Association of College and Research Libraries (ACRL, 2011). La taxonomía de la ACRL proporciona la base epistemológica adecuada para articular el Diseño Centrado en Evidencias que estructura la prueba: sus siete estándares —que van desde la localización y evaluación crítica de imágenes hasta su uso ético y la creación autónoma (Hattwig et al., 2013)— se convierten en las Afirmaciones del DCE, permitiendo descomponer competencias complejas en conductas observables y medibles. El detalle de cada estándar con sus indicadores de desempeño y resultados de aprendizaje se presenta en el Anexo A.

En 2022, la ACRL publicó un nuevo marco de alfabetización visual alineado con los Estándares de competencia de alfabetización visual dela ACRL para la educación superior de 2011

(ACRL, 2022). El documento reorganiza el campo en cuatro grandes temas —participación en el panorama visual cambiante, percepción de imágenes como información, discernimiento crítico y justicia social en la práctica visual— y reemplaza la lógica de estándares con indicadores medibles por una lógica de prácticas y disposiciones de desarrollo continuo y no jerárquico. La actualización enriquece considerablemente la comprensión del campo al incorporar fenómenos contemporáneos como algoritmos, deepfakes, inteligencia artificial y dimensiones de equidad e inclusión. Esa misma reorientación, sin embargo, la hace incompatible con los requerimientos del Diseño Centrado en Evidencias adoptado en esta investigación, que exige indicadores de desempeño concretos y resultados de aprendizaje verificables para la construcción de ítems cerrados. Por esa razón, el presente estudio retiene la taxonomía de 2011 como base operativa del instrumento; la perspectiva crítica del marco de 2022 ofrece un horizonte relevante para versiones futuras que incorporen dimensiones de lectura crítica en entornos digitales. Investigaciones recientes confirman que el campo está en transformación activa: Brumberger (2025) y Thomson et al. (2025) documentan cómo la IA generativa desplaza las fronteras entre lectura y escritura visual y exige nuevas competencias críticas, lo que refuerza la pertinencia de versiones futuras del instrumento que incorporen esas dimensiones.

### 2.2. Evolución y Brecha en la Evaluación Objetiva

Los estándares de la ACRL definen qué debe saber hacer un estudiante visualmente alfabetizado, pero no dicen cómo medirlo. Ver Anexo A. Esa distancia entre la definición y la medición es la brecha que esta investigación intenta reducir. La evaluación de estas competencias ha seguido históricamente dos caminos divergentes: el del juicio experto —portafolios, crítica grupal, rúbricas institucionales— y el de los instrumentos estandarizados con fundamento psicométrico. El primero, señala Kędra (2018), tiende a concentrarse en la pericia técnica o el

talento artístico, dejando en segundo plano las competencias crítico-visuales que son precisamente las más difíciles de evaluar: las habilidades hermenéuticas, el razonamiento ético y la decodificación analítica. El segundo camino, más reciente, propone reemplazar esa subjetividad por criterios psicométricos verificables. Es desde ese segundo camino desde donde se construye este trabajo.

La trayectoria de los instrumentos objetivos para medir alfabetización visual muestra una búsqueda constante por equilibrar la amplitud del constructo (Nunnally y Bernstein, 1994) con la eficiencia práctica. El aporte más influyente fue el *Visual Literacy Assessment Test* (VLAT), desarrollado por Lee et al. (2016, 2017): una batería de 53 ítems diseñada para evaluar la capacidad de interpretar visualizaciones de datos y gráficos complejos. El VLAT estableció un estándar metodológico, pero su extensión resultó problemática en contextos educativos reales. Aplicar 53 reactivos genera fatiga cognitiva y desmotivación, factores que introducen ruido estadístico y amenazan la validez de los puntajes obtenidos.

El paso siguiente fue reducir esa carga sin perder precisión. Cui et al. (2023) desarrollaron versiones adaptativas del VLAT —el A-VLAT y el A-CALVI— basadas en la Teoría de Respuesta al Ítem en lugar de la Teoría Clásica. Al calibrar los bancos de preguntas bajo la lógica probabilística de la TRI y aprovechar la invarianza[3] de los parámetros de dificultad y discriminación (Muñiz, 2010), lograron seleccionar en cada aplicación solo los ítems más informativos para el nivel de habilidad del estudiante evaluado. El resultado fue notable: redujeron

[3] Invarianza de la dificultad: Dificultad real del ítem estimada de manera independiente a la muestra de estudiantes utilizada para la calibración.
Invarianza de la discriminación: Capacidad de un ítem para diferenciar entre sujetos con distintos niveles de habilidad (parámetro de discriminación) se mantiene constante independientemente del grupo evaluado. Adaptado de Muñiz, 2010.

la prueba de 53 a un promedio de 27 ítems sin comprometer la confiabilidad, que se mantuvo por encima de 0.90 en todas las versiones.

Pandey y Ottley (2023) confirmaron esta dirección con el Mini-VLAT, demostrando que los formatos abreviados son herramientas diagnósticas eficaces cuando el tiempo de aplicación es limitado. Estos hallazgos justifican directamente el diseño del presente prototipo: un instrumento de 34 ítems que aplica primero los filtros de la Teoría Clásica y avanza luego hacia la calibración bajo el Modelo de Rasch, buscando el mismo equilibrio entre precisión diagnóstica y viabilidad práctica que las investigaciones recientes han demostrado posible.

### 2.3. El Diseño Centrado en Evidencias (DCE) para la construcción de la prueba

Los estándares de la ACRL describen competencias en términos abstractos: interpretar, evaluar, crear, usar éticamente. Para convertir esas declaraciones en ítems de una prueba objetiva hace falta un procedimiento que explicite qué conducta observable demuestra cada competencia y bajo qué condiciones. Ese procedimiento es el Diseño Centrado en Evidencias (DCE), propuesto por Mislevy, Steinberg y Almond (2003). A diferencia de los enfoques tradicionales, que suelen partir de la redacción intuitiva de preguntas sobre el temario, el DCE trata el diseño de una prueba como la construcción de un argumento probatorio: cada ítem debe poder justificarse como evidencia válida de que el estudiante posee la competencia que se declara medir.

Este paradigma ha demostrado tal solidez que ha sido adoptado por sistemas de evaluación de alto impacto en varios países. En Colombia, el ICFES lo incorporó en 2007 para estructurar las pruebas Saber Pro, que evalúan anualmente a los estudiantes universitarios antes de graduarse (ICFES, 2018). El modelo operacionaliza el DCE en tres capas interdependientes: el Modelo del Estudiante (qué competencias se mide), el Modelo de Evidencias (qué conductas demuestran esas competencias) y el Modelo de Tareas (los estímulos concretos que provocan esas conductas). Esta

investigación sigue esa misma lógica, homologando las tres capas con la estructura de los estándares de la ACRL:

Las Afirmaciones — Modelo del Estudiante Se derivan de cada uno de los siete estándares de la ACRL y constituyen declaraciones sobre los conocimientos y habilidades que se espera que el estudiante posea. Por ejemplo: "El estudiante interpreta y analiza críticamente los significados implícitos y explícitos de las imágenes y los medios visuales". Las afirmaciones anclan el instrumento al constructo central y garantizan que ningún ítem quede sin respaldo teórico.

Las Evidencias — Modelo de Evidencias Se derivan de los Indicadores de Desempeño de la ACRL y traducen cada afirmación abstracta en comportamientos concretos y verificables.

Las Tareas — Modelo de Tareas Se derivan de los Resultados de Aprendizaje de la ACRL y se materializan en los ítems de selección múltiple con única respuesta (SMUR). Cada tarea se diseñó para presentar un estímulo visual controlado y opciones de respuesta que obliguen al estudiante a manifestar la evidencia requerida.

En el ensamblaje final de los ítems, Cui et al. (2023) advierten sobre la necesidad de aplicar lo que denominan "restricciones no psicométricas": la selección de reactivos no debe depender solo de su comportamiento estadístico, sino garantizar también la representatividad de todas las dimensiones teóricas de la ACRL. Sin esta restricción, un instrumento puede volverse estadísticamente impecable, pero perder validez de contenido al omitir competencias críticas difíciles de medir, como el uso ético de la información visual.

**Tabla 1.**

*Homologación del marco de especificaciones del DCE con la estructura de los estándares de la ACRL.*

| DISEÑO CENTRADO EN EVIDENCIAS | ESTÁNDARES DE ALFABETIZACIÓN VISUAL ACRL |
|---|---|

| | |
|---|---|
| Afirmaciones | Estándares |
| Evidencias | Indicadores de desempeño |
| Tareas | Resultados de aprendizaje |

Nota: Elaboración propia.

## 2.4. Marco Psicométrico: Teoría Clásica y Teoría de Respuesta al Ítem

Con la estructura lógica del instrumento consolidada mediante el DCE, el siguiente paso fue elegir el modelo matemático para analizar y validar empíricamente los datos del piloto. La evaluación estandarizada contemporánea se rige por la Teoría de Respuesta al Ítem (TRI) y, más específicamente, por el Modelo de Rasch (Bond & Fox, 2015). Sin embargo, aplicar directamente estos modelos probabilísticos exige matrices de datos depuradas y el cumplimiento previo de supuestos como la unidimensionalidad del constructo. Por eso, esta investigación adoptó una estrategia secuencial: primero un tamizaje bajo la Teoría Clásica de los Tests (TCT) y luego la calibración paramétrica bajo el Modelo de Rasch.

El tamizaje clásico se apoya en dos parámetros fundamentales: índice de dificultad (p) y correlación punto-biserial (rpbis). El primero indica la proporción de estudiantes que respondió correctamente cada ítem. La correlación punto-biserial (rpbis) mide la capacidad de cada reactivo para distinguir a los estudiantes con mayor competencia visual de los que tienen menor competencia (Muñiz, 2010). Aunque la TCT tiene limitaciones estructurales —las propiedades de los ítems dependen de la muestra con que se calculan—, sus estadísticos permiten detectar tempranamente reactivos defectuosos: aquellos con discriminación negativa o cercana a cero que contaminarían los análisis posteriores. Depurar esos ítems antes de proceder a la calibración TRI es un requisito metodológico ineludible.

La TRI y el Modelo de Rasch ofrecen ventajas sobre la TCT que resultan especialmente relevantes para los propósitos de esta investigación. La más importante para su viabilidad institucional es la invarianza de los parámetros: la dificultad de cada ítem y la habilidad de cada

estudiante se estiman con independencia de la muestra específica evaluada, lo que hace posible comparar resultados entre cohortes distintas y construir bancos de ítems escalables. El Modelo de Rasch añade a esto la capacidad de ubicar la habilidad de las personas y la dificultad de los ítems en una misma unidad de medida —el logit—, generando una escala de intervalo que hace posible generar el Mapa de Wright como una representación gráfica que permite diagnosticar si el instrumento está bien calibrado para la población objetivo. A diferencia de la TCT, que asume un error de medida constante para todos los puntajes, el modelo reconoce que la precisión varía según la distancia entre la habilidad del estudiante y la dificultad del ítem, lo que permite identificar con exactitud en qué rangos de la escala hacen falta reactivos adicionales.

Estas propiedades también sustentan el horizonte adaptativo del instrumento. Como señalan Cui et al. (2023), la invarianza de los parámetros de discriminación y dificultad permite construir bancos de ítems dinámicos en que la precisión diagnóstica no depende de la longitud fija de la prueba sino de la calidad informativa de cada reactivo seleccionado para el nivel específico del estudiante, abriendo la puerta a futuras versiones adaptativas del instrumento.

La validación del instrumento se concibe, siguiendo a Kane (2013), como un argumento interpretativo que exige múltiples fuentes de evidencia articuladas mediante triangulación metodológica (Denzin, 2009): el juicio experto garantiza la relevancia teórica del contenido, mientras el análisis empírico revela el comportamiento real de los ítems, y la convergencia o divergencia entre ambas fuentes orienta los ajustes necesarios (Cohen, Manion & Morrison, 2018).

Los pilares descritos en este capítulo configuran el andamiaje conceptual del instrumento. El capítulo siguiente describe cómo ese andamiaje se convirtió en decisiones concretas de diseño, validación y análisis.

# CAPÍTULO 3. METODOLOGÍA

Este capítulo describe las decisiones metodológicas que permitieron convertir el marco conceptual en un instrumento concreto: el enfoque de investigación, la delimitación de la población, el proceso de construcción del instrumento y los procedimientos de validación y análisis psicométrico adoptados.

## 3.1. Enfoque y Diseño de la Investigación

Esta investigación adopta un enfoque cuantitativo de tipo instrumental, orientado al diseño y validación de propiedades psicométricas de un instrumento de medición (Ato et al., 2013). La elección responde a la naturaleza del problema: determinar si un instrumento mide lo que declara medir exige indicadores numéricos de dificultad, discriminación y confiabilidad —índices que la metodología cualitativa no puede producir (Hernández-Sampieri et al., 2014; Nunnally y Bernstein, 1994). Este enfoque facilita la obtención de indicadores numéricos precisos y métricas de error estandarizadas para evaluar la calidad técnica de la prueba, incluyendo la estimación de índices de dificultad, capacidad de discriminación de los ítems y confiabilidad de la escala.

En la psicometría contemporánea se establece que la validez es un argumento interpretativo: validar una prueba implica reunir múltiples fuentes de evidencia para justificar que las inferencias derivadas de sus puntajes son legítimas y apropiadas para el contexto en que se usan. Kane (2013). Desde esa perspectiva, este estudio adopta un diseño de validación convergente que articula tres tipos de evidencia: la alineación de los ítems con el dominio teórico de la ACRL (evidencia sustantiva), el juicio de pares académicos (evidencia experta) y el comportamiento empírico de los estudiantes frente a los ítems (evidencia estructural). Su desarrollo se presenta en el Capítulo 4.

### 3.2. Contexto y Población de Estudio

El instrumento se aplicó en el programa de Diseño Digital y Multimedia de la Universidad Colegio Mayor de Cundinamarca, institución estatal ubicada en Bogotá. La elección de este programa no es arbitraria pues corresponde de manera directa con la naturaleza epistémica de la alfabetización visual, cuyo perfil de egreso declara competencias de representación y comunicación visual compatibles con el constructo de alfabetización visual (ver Capítulo 1).

La población objetivo se delimitó con un criterio de inclusión preciso. Se seleccionaron únicamente estudiantes activos de noveno semestre que hubieran cursado y aprobado al menos el 80% de los créditos del plan de estudios. Evaluar a estudiantes en la etapa culminante de su formación permite contrastar el instrumento con el nivel más alto de competencia visual que el programa declara desarrollar, lo que resulta indispensable para calibrar los niveles de dificultad del test frente al perfil de egreso institucional.

Para el piloto se utilizó un muestreo no probabilístico por conveniencia, con una muestra de n=54 estudiantes. En esta fase se buscaba examinar el comportamiento técnico de los ítems para detectar anomalías, establecer los estimadores de dificultad y verificar la unidimensionalidad del constructo como requisito previo al Modelo de Rasch.

El tamaño de la muestra está respaldado por la literatura especializada. Hogarty et al. (2005) y Muñiz (2010) señalan que, en fases de tamizaje, muestras de entre 50 y 100 participantes son suficientes para estabilizar los estimadores básicos de dificultad y consistencia interna y para detectar anomalías graves como la discriminación negativa. Linacre (1994) es más preciso: con 50 individuos es posible estimar la dificultad de los ítems con un nivel de confianza del 99% y un margen de error de ±1 logit. Estos requerimientos son distintos de los que exige una calibración paramétrica definitiva bajo la TRI, que necesita muestras superiores a 200 casos para garantizar la

estabilidad de los parámetros de dificultad y discriminación con independencia de la muestra. Dado que el alcance de esta investigación es la validación del prototipo y no su calibración definitiva, n=54 satisface los requerimientos técnicos de esta fase y establece las condiciones empíricas para la siguiente.

## 3.3. Diseño y construcción del Instrumento

La construcción del instrumento fue el proceso más exigente de la investigación. El reto central fue garantizar que cada ítem midiera exactamente la competencia prevista, y no el conocimiento de un idioma, la familiaridad con plataformas específicas o la capacidad de leer enunciados ambiguos. Las secciones siguientes describen cómo se abordó ese problema en cuatro versiones sucesivas del instrumento, apoyadas en el Diseño Centrado en Evidencias y en herramientas de Inteligencia Artificial Generativa.

### 3.3.1. Especificación del modelo bajo el Diseño Centrado en Evidencias

El punto de partida fue traducir la taxonomía de la ACRL (2011) en comportamientos observables y medibles mediante un formato de evaluación cerrado, siguiendo la lógica del DCE presentada en la sección 2.3.

Cada uno de los siete estándares de la ACRL se formuló como una afirmación directa sobre lo que el estudiante debe saber hacer. A partir de cada afirmación, los Indicadores de Desempeño de la ACRL permitieron desagregarla en conductas observables y verificables (Anexo B). Cada conducta se materializó, a su vez, en un ítem de Selección Múltiple con Única Respuesta (SMUR), derivado directamente de los Resultados de Aprendizaje de la ACRL. Este formato se eligió por su eficiencia para el procesamiento estadístico y su objetividad en la calificación. El principio rector en la construcción de cada ítem fue aislar la variable a medir garantizando que los distractores no fueran opciones arbitrarias sino representaciones de errores cognitivos, técnicos o

perceptivos frecuentes en estudiantes de diseño, lo que asegura que un acierto o un error refleje el nivel de competencia visual del evaluado y no deficiencias de redacción ni ambigüedades en el estímulo. El detalle completo de afirmaciones, evidencias y tareas para cada uno de los siete estándares se incluye en el Anexo B.

### 3.3.2. Proceso iterativo de construcción del instrumento

La taxonomía de la ACRL está estructurada en 7 estándares (Afirmaciones), 24 indicadores de desempeño (Evidencias) y 115 resultados de aprendizaje (Tareas).

El primer paso fue determinar cuáles de esos 115 resultados de aprendizaje podían evaluarse mediante una prueba objetiva. El análisis inicial clasificó 36 como prioritarios, 39 como opcionales y 40 como no evaluables en formato cerrado. El resultado de este análisis se incluye en el Anexo B. Matriz de Evaluabilidad de Resultados de Aprendizaje.

La construcción del banco de ítems se desarrolló en cuatro versiones sucesivas, cada una con varias iteraciones internas. Desde el inicio surgió un desafío técnico propio de las pruebas de competencias visuales: las imágenes de repositorios web introducen sesgos culturales, ruido visual no controlado y problemas de derechos de autor. Para resolverlo, la investigación recurrió a herramientas de Inteligencia Artificial Generativa, que permitieron ejercer un control preciso sobre las variables del estímulo, proceso que es desarrollado normalmente por equipos interdisciplinarios en la construcción de pruebas estandarizadas a gran escala.

Versión 1 (Fase exploratoria): se generó una batería preliminar de 45 ítems —los 36 prioritarios más 9 opcionales— usando ChatGPT-4 como asistencia técnica para explorar posibles escenarios de evaluación dentro del marco DCE. La revisión interna detectó tres problemas: precisión insuficiente en las tareas a medir, insuficiencia de criterios psicométricos especializados y redundancia en la medición de habilidades de bajo nivel taxonómico.

Versión 2 (Fase de refinamiento técnico): el instrumento se depuró y redujo a 36 ítems, agrupados en tres bloques temáticos siguiendo la estructura de Kędra (2018):

a) Bloque 1. Determinación y búsqueda de imágenes: capacidad para definir cuándo son necesarias las imágenes y localizarlas, 10 ítems. b) Bloque 2. Lectura y escritura visual: capacidad para interpretar y comunicarse mediante imágenes, 21 ítems. c) Bloque 3. Ética visual: capacidad para utilizar imágenes de manera responsable, 3 ítems. Este bloque quedó reducido por las dificultades de evaluar esta competencia en formato cerrado.

Los niveles de dificultad se distribuyeron según la Taxonomía de Bloom en tres grupos equivalentes: 12 ítems de nivel alto, 12 de nivel medio y 12 de nivel bajo.

Cada ítem se redactó desde cero como borrador técnico completo, con los siguientes componentes: código, bloque temático, afirmación, evidencia, tarea, propósito, tipo de estímulo, nivel cognitivo, nivel de dificultad, modelo psicométrico, tipo de estímulo, descripción del estímulo, enunciado, opciones de respuesta, justificación de la clave correcta, análisis de distractores y observaciones técnicas. Los estímulos visuales se describieron en texto con las indicaciones necesarias para su creación posterior. El borrador técnico completo se incluye en el Anexo C. Especificaciones Técnicas del Instrumento de Medición.

El principal desafío de esta versión fue controlar la varianza irrelevante, para garantizar que la dificultad de cada ítem proviniera de la complejidad de la competencia visual y no de ambigüedades semánticas en los enunciados, de la calidad de las imágenes, del conocimiento de otro idioma o de distractores implausibles que permitieran acertar por descarte. Ajustar las opciones de respuesta para que todos los distractores fueran funcionales y atractivos para estudiantes de menor habilidad exigió múltiples iteraciones. Al final de este proceso se

reformularon completamente 8 ítems y se eliminaron 2 por no alcanzar el nivel de calidad requerido.

Versión 3 (Consolidación de reactivos): con 34 ítems distribuidos entre los tres niveles de complejidad de Bloom, el trabajo central de esta fase fue la creación y selección de los estímulos visuales. Los ítems exigían que el estudiante evaluara problemas específicos de composición, color, jerarquía visual o manipulación de imágenes, lo que hacía inviable usar imágenes de repositorios web (stock), pues las imágenes de stock pueden contener elementos visuales ajenos a lo que el ítem busca medir. Los motores de IA generativa resolvieron ese problema, permitiendo producir imágenes diseñadas con precisión que incorporaran exactamente el error técnico, la distorsión semántica o la composición específica que cada ítem del DCE requería, sin ruido visual adicional.

Esta versión se sometió a una valoración preliminar confidencial con 5 estudiantes de noveno semestre que no participarían en el piloto. Sus observaciones permitieron afinar la terminología técnica de algunos ítems, corregir errores gramaticales que habían pasado inadvertidos y mejorar la calidad de las imágenes.

Versión 4 (Prototipo final): corregidas las deficiencias de la tercera versión, se construyeron los tres instrumentos derivados: el formulario del piloto en Google Forms, el cuadernillo de instrucciones para jueces y el formulario de evaluación para jueces. El cuadernillo de 40 páginas describía la estructura de la prueba, el proceso de validación, los criterios de evaluación (claridad, coherencia, suficiencia, relevancia) y la guía de calificación (ver Anexo E).

La prueba final constaba de 34 ítems en los tres bloques temáticos definidos desde la Versión 2. Esa longitud responde a un criterio de eficiencia respaldado por Cui et al. (2023), quienes demostraron que es posible reducir pruebas de 53 ítems a versiones de 27 sin comprometer

la precisión psicométrica, manteniendo confiabilidad superior a 0.90. Un instrumento de 34 ítems bien seleccionados maximiza la información diagnóstica por reactivo y reduce la fatiga de los participantes sin sacrificar la cobertura del constructo.

### 3.3.3. Uso de inteligencia artificial

A lo largo del proceso de investigación se utilizaron herramientas de IA con tres propósitos distintos: gestión documental, construcción de ítems y creación de estímulos visuales.

#### *3.3.3.1. Investigación y gestión documental*

Para la revisión bibliográfica, el análisis de referencias y la redacción inicial de los reactivos de las versiones 1 y 2 se trabajó con ChatGPT 4.0 y 5.0, organizando el trabajo en cuatro proyectos diferenciados dentro de la plataforma. La revisión del estado del arte se apoyó adicionalmente en Scispace AI y en Gemini, que tras la aplicación del piloto se utilizó también en el análisis estadístico preliminar y la estructuración inicial del informe. Sin embargo, como los resultados generados por estas herramientas no cumplían completamente las exigencias técnicas de la construcción de ítems psicométricos, ya que las iteraciones sucesivas generaban confusiones y errores acumulados, cada reactivo fue revisado y ajustado manualmente. Las correcciones que condujeron a la tercera y cuarta versión del instrumento se realizaron íntegramente sin asistencia de IA.

Nota sobre el proceso de escritura: la revisión de estilo y los ajustes de redacción de la versión final de este documento contaron con el apoyo del modelo Claude Sonnet 4.6 de Anthropic. Su función se limitó estrictamente a la corrección de expresión escrita. El diseño del estudio, el análisis de los datos y las conclusiones son de responsabilidad exclusiva del autor.

#### *3.3.3.2. Creación de estímulos visuales*

Para generar las imágenes de los reactivos con estímulos visuales se utilizaron DALL·E/ChatGPT Imagen, Qwen, Nano Banana y Deevid AI, entre otras herramientas. Mediante instrucciones estructuradas bajo parámetros de diseño precisos, se produjeron 142 imágenes, de las cuales 24 se incluyeron en la prueba piloto final.. Ver Anexo D. Protocolo y registro de trabajo documental y generación de estímulos visuales mediante Inteligencia Artificial.

El recurso a la IA generativa para la producción de estímulos visuales implica algunas limitaciones que es necesario explicitar. Los modelos de difusión entrenados con grandes conjuntos de datos tienden a reproducir patrones estéticos dominantes en sus corpus de entrenamiento, lo que puede introducir sesgos de homogeneidad estilística —una recurrencia de composiciones, paletas cromáticas y convenciones visuales propias de la fotografía comercial occidental— y subrepresentar tradiciones gráficas de otros contextos culturales. Adicionalmente, la calidad del estímulo generado depende de la precisión de las instrucciones textuales proporcionadas al modelo, lo que traslada al investigador la responsabilidad de especificar con rigor las variables visuales requeridas por cada ítem. Para mitigar estos riesgos, las 142 imágenes generadas fueron sometidas a un proceso de revisión en tres niveles consecutivos: edición técnica con Adobe Photoshop e Illustrator para corregir deficiencias de resolución y composición, revisión del investigador para verificar la correspondencia entre el estímulo y las exigencias del DCE, y valoración preliminar con cinco estudiantes de noveno semestre que permitió detectar ambigüedades visuales no previstas. El panel de jueces expertos evaluó posteriormente la suficiencia y claridad de cada estímulo visual como parte integral de la validación de contenido.

### 3.4. Validación de Contenido por Juicio de Expertos

Que un ítem trate el tema correcto no garantiza que lo mida correctamente. Para verificar que los reactivos representaban con fidelidad las competencias definidas por los estándares ACRL y estructuradas mediante el DCE, el prototipo fue evaluado por un panel de seis jueces expertos.

Los criterios de selección del panel fueron simultáneos y no negociables: título de posgrado (maestría o superior), experiencia docente en educación superior de al menos 15 años, y trayectoria investigativa o profesional activa en diseño digital, comunicación visual o evaluación educativa. Este perfil interdisciplinar buscó garantizar que la revisión cubriera tanto la corrección técnica de los conceptos visuales como su pertinencia pedagógica y su viabilidad evaluativa. Cada juez recibió el cuadernillo de instrucciones y tareas (ver Anexo E. Cuadernillo de validación para panel de jueces expertos).

El procedimiento se sistematizó mediante una matriz de evaluación con escala Likert de cuatro niveles (1 = Deficiente, 4 = Sobresaliente). La escala de número par se eligió deliberadamente para forzar un posicionamiento direccional por parte del juez, evitando la generación de una tendencia central. Cada juez evaluó de forma independiente los 34 ítems en cuatro dimensiones:

1. Claridad: si la sintaxis del enunciado y la calidad del estímulo son comprensibles y libres de ambigüedades.

2. Coherencia: si hay alineación lógica entre la tarea exigida por el ítem y el Resultado de Aprendizaje asociado.

3. Suficiencia: si la información provista —texto e imagen— es completa para que un estudiante competente deduzca la respuesta sin suposiciones externas.

4. Relevancia: si el reactivo es pertinente y necesario para medir la competencia de alfabetización visual asociada al Resultado de Aprendizaje específico.

Cada ítem incluía además un campo opcional para observaciones.

Para cuantificar el consenso entre jueces se descartaron los porcentajes simples de acuerdo por su vulnerabilidad estadística ante el azar, y se empleó el Coeficiente V de Aiken con un umbral de aceptación de $V \geq 0.80$ (Escobar-Pérez y Cuervo-Martínez, 2008). Los ítems con V inferior a ese umbral en Relevancia o Coherencia se consideraron candidatos a eliminación o reformulación sustancial; valores bajos en Claridad se interpretaron como señal de alerta para ajustes de redacción. Complementariamente, el Coeficiente W de Kendall permitió medir el grado de concordancia global entre jueces e identificar ambigüedades y discrepancias en la interpretación de los ítems.

### 3.5. Aplicación del Piloto

La prueba se implementó en modalidad digital asincrónica mediante un formulario de Google Forms. Esta plataforma permitió presentar las imágenes en su resolución nativa, garantizando condiciones técnicas homogéneas para todos los participantes. Junto con las respuestas, se registraron los tiempos de ejecución total de la prueba.

Para cada uno de los 34 ítems se recolectaron dos tipos de datos. El primero fue la respuesta cognitiva: la selección de una opción en el formato de selección múltiple con única respuesta. El segundo fue la respuesta metacognitiva: cada estudiante valoró, mediante escalas estandarizadas, su percepción sobre la dificultad del ítem y la claridad del enunciado, las instrucciones y los elementos visuales. Estos datos periféricos se cruzaron posteriormente con los hallazgos del análisis psicométrico.

La aplicación cumplió con los estándares éticos para la investigación en educación. Antes de acceder a los ítems, los participantes leyeron y aceptaron un Consentimiento Informado que explicaba la naturaleza instrumental del estudio, garantizaba el anonimato en el tratamiento de los datos y aclaraba que los resultados no tendrían ninguna incidencia sobre sus calificaciones académicas. Esta última condición es metodológicamente relevante: reducir la presión académica sobre el estudiante disminuye el riesgo de respuestas aleatorias y mejora la calidad de los datos que entran al análisis psicométrico.

## 3.6. Estrategia de Análisis Psicométrico

Los datos del piloto se procesaron con el software JAMOVI (versión 2.7.15), usando los módulos *snowIRT* y *Rasch Models*. El análisis siguió una estrategia secuencial en dos fases: primero un tamizaje bajo la Teoría Clásica de los Tests (TCT) y luego la calibración paramétrica bajo el Modelo de Rasch.

### 3.6.1. Fase de Tamizaje (Teoría Clásica de los Tests)

La TCT tiene limitaciones estructurales conocidas, pues sus estimaciones de dificultad y habilidad son mutuamente dependientes de la muestra con que se calculan. Sin embargo, sus estadísticos siguen siendo herramientas útiles en fases exploratorias para detectar anomalías graves antes de proceder a análisis más complejos (Muñiz, 2010). En esta fase, siguiendo las recomendaciones de Nunnally y Bernstein (1994), la matriz de datos se sometió a la estimación de tres parámetros:

1. Índice de Dificultad (p): proporción de aciertos por ítem. Permitió detectar efectos de techo —ítems que casi todos respondieron correctamente— y efectos de suelo —ítems tan complejos o confusos que la tasa de acierto no superó el azar.

2. Índice de Discriminación (rpbis): la correlación punto-biserial evaluó la consistencia interna de cada reactivo frente al puntaje global. Una correlación negativa es señal de un ítem defectuoso: indica que los estudiantes con menor competencia visual lo aciertan más que los competentes. Esos ítems se marcaron para revisión cualitativa o exclusión antes del análisis psicométrico basado en la TRI (Muñiz, 2010).

3. Confiabilidad Global (KR-20): el coeficiente Kuder-Richardson 20 —equivalente dicotómico del Alfa de Cronbach— estimó la consistencia interna preliminar del instrumento y verificó su homogeneidad antes de la reducción dimensional.

### 3.6.2. Fase de Calibración (Modelo de Rasch)

Con la matriz depurada, los datos de los 32 ítems retenidos se sometieron a calibración bajo el Modelo de Rasch. La ejecución contempló tres procedimientos analíticos secuenciales (Bond & Fox, 2015):

1. Análisis de Dimensionalidad: el Modelo de Rasch exige que todos los ítems midan un único rasgo latente, lo cual verificó con el análisis de componentes principales sobre los residuales estandarizados —la diferencia entre las respuestas observadas y las predicciones del modelo—, determinando si la varianza residual es ruido aleatorio o si esconde una segunda dimensión que esté contaminando la medición del constructo principal.

2. Estimación de Parámetros de Ajuste: se calcularon los estadísticos de ajuste interno (*Infit Mean Square*) y externo (*Outfit Mean Square*) para cada ítem y cada estudiante. Estos valores cuantifican la discrepancia entre las respuestas observadas y las esperadas por el modelo (Bond & Fox, 2015). Se fijaron rangos de tolerancia entre 0.5 y 1.5 logits; los reactivos fuera de ese rango se clasificaron como erráticos o redundantes por introducir varianza irrelevante para el constructo.

3. Construcción del Mapa de Wright: aprovechando la escala conjunta del Modelo de Rasch, que ubica la habilidad de las personas y la dificultad de los ítems en la misma unidad métrica, se generó el Mapa de Wright, visualizando simultáneamente la distribución jerárquica de la dificultad de los ítems y la distribución del nivel de competencia de los estudiantes, ofreciendo un diagnóstico directo sobre la precisión del instrumento para esta población.

Con el prototipo construido, el contenido validado por jueces, los datos del piloto procesados mediante TCT y el Modelo de Rasch; y articulando la evidencia cualitativa y cuantitativa mediante la estrategia de triangulación convergente, el capítulo siguiente presenta los hallazgos y discute su significado para la validez del instrumento y para la formación en diseño.

# CAPÍTULO 4. ANÁLISIS Y DISCUSIÓN DE RESULTADOS

Validar un instrumento psicométrico es un procedimiento de verificación que en los términos de Kane (2013), implica la construcción progresiva de un argumento interpretativo. Desde esa perspectiva, este capítulo desarrolla ese argumento a partir de dos fuentes de datos recogidas de manera simultánea: el juicio de seis expertos sobre el contenido del instrumento y el desempeño de 54 estudiantes de noveno semestre durante la aplicación piloto.

El análisis se organiza en torno a tres preguntas de fondo: si los ítems representan adecuadamente el constructo de alfabetización visual, si funcionan con consistencia psicométrica cuando los responden estudiantes reales, y si el instrumento cumple las condiciones técnicas para ser calibrado bajo el Modelo de Rasch en una fase posterior. Para responderlas, el capítulo examina primero la validez de contenido evaluada por el panel de jueces (sección 4.2); a continuación, la estructura interna del instrumento —dificultad, discriminación, dimensionalidad y confiabilidad bajo la Teoría Clásica de los Tests— derivada de los datos del piloto (sección 4.3); luego, la calibración Rasch y el diagnóstico conjunto persona-ítem a través del Mapa de Wright (sección 4.4); después, el contraste entre la dificultad teórica planteada en el diseño y la dificultad observada empíricamente (sección 4.5); y, finalmente, la triangulación de ambas fuentes de evidencia con sus convergencias, divergencias e implicaciones para el desarrollo del instrumento (sección 4.6).

## 4.1. Introducción a la Validación Convergente

El diseño de validación adoptado ejecutó de forma simultánea la evaluación por jueces y la aplicación piloto, procedimientos que habitualmente se aplican de manera secuencial. Al recabar al mismo tiempo la evidencia teórica —lo que los expertos consideran que cada ítem mide— y la evidencia empírica —lo que el comportamiento estadístico de los estudiantes revela que ese ítem efectivamente produce—, fue posible contrastar directamente las hipótesis de validez con su

funcionamiento real. Cuando ambas fuentes señalan en la misma dirección, la evidencia gana solidez. Cuando apuntan en sentidos distintos, el contraste mismo adquiere valor diagnóstico.

Esta aproximación responde al principio de triangulación metodológica propuesto por Denzin (2009) y desarrollado por Cohen, Manion y Morrison (2018), quienes sostienen que el uso concurrente de múltiples fuentes fortalece la interpretación al compensar las limitaciones propias de cada fuente individual. El juicio experto aporta una lectura teórica capaz de detectar problemas de redacción y pertinencia que el análisis estadístico difícilmente anticipa. Los datos empíricos, al mostrar cómo responden realmente los estudiantes, completan el diagnóstico con información que ningún juicio a priori puede suministrar.

Dado el tamaño de la muestra —característico de una fase exploratoria—, el análisis cuantitativo se apoyó en los indicadores de la Teoría Clásica de los Tests, fundamentalmente el índice de dificultad, la capacidad de discriminación y la consistencia interna. Muñiz (2010) documenta su utilidad en estas etapas iniciales para identificar anomalías severas en el banco de ítems antes de proceder a calibraciones paramétricas más exigentes. Sus resultados cumplen también una función adicional relevante para este estudio. La consistencia interna, además de estimar la fiabilidad de la prueba, opera como indicador preliminar de unidimensionalidad, condición que el Modelo de Rasch exige antes de la calibración definitiva. Solo cuando los reactivos pueden interpretarse como medidas de un único rasgo latente dominante es posible obtener la propiedad de invarianza que caracteriza al modelo, gracias a la cual la estimación de la habilidad del estudiante resulta independiente de los ítems específicos que respondió, y la dificultad de cada reactivo permanece estable frente a muestras distintas (Bond & Fox, 2015).

La integración de una estructura de contenido validada por expertos con un comportamiento estadístico funcional confirmado por el piloto constituye la base técnica que justifica avanzar hacia una medición psicométrica estandarizada del instrumento.

### 4.2. Análisis de Validez de Contenido

El banco de 34 reactivos fue sometido al juicio de seis expertos, quienes evaluaron cada ítem según los cuatro criterios de claridad, coherencia, suficiencia y relevancia. Los resultados se cuantificaron mediante el coeficiente V de Aiken (Escobar-Pérez y Cuervo-Martínez, 2008), con un umbral de aceptación de $V \geq 0.80$. Para la interpretación de los resultados se estableció una distinción operativa entre los criterios. Los valores por debajo del umbral en Relevancia o Coherencia indicarían que el ítem no mide lo que pretende y requeriría eliminación o reformulación, mientras que valores bajos en Claridad se tratarían como señal de alerta para ajustes de redacción, sin implicar necesariamente la eliminación del contenido.

Los resultados globales fueron satisfactorios en las dimensiones centrales del constructo. Relevancia ($V = 0.94$) y Coherencia ($V = 0.93$) obtuvieron los promedios más altos, confirmando que la operacionalización de los estándares de la ACRL mediante el DCE es pertinente para el contexto de la educación superior en diseño. El consenso de los jueces en estas dos dimensiones es significativo: las situaciones evaluativas propuestas —búsqueda, interpretación y uso ético de imágenes— fueron reconocidas como representativas de las demandas reales del ejercicio profesional en el perfil del diseñador digital y corresponden con el conjunto de competencias visuales requeridas en estudiantes próximos a graduarse. La Tabla 2, presentada a continuación, detalla los valores del coeficiente V de Aiken para cada uno de los 34 ítems. Los valores por debajo del umbral de 0.80 aparecen en negrilla.

**Tabla 1.**

*Valores del coeficiente V de Aiken por ítem y criterio de evaluación.*

| Ítems | Promedio ítem en Relevancia | Promedio ítem en Coherencia | Promedio ítem en Suficiencia | Promedio ítem en Claridad |
|---|---|---|---|---|
| **Promedio global por criterio** | *0.94* | *0.93* | *0.81* | *0.80* |
| **1** | 0.94 | 0.89 | **0.67** | **0.67** |
| **2** | 0.94 | 0.94 | 0.89 | 0.89 |
| **3** | 0.89 | 0.94 | 0.89 | 0.89 |
| **4** | 1.00 | 1.00 | 0.83 | 0.83 |
| **5** | 1.00 | 0.89 | **0.67** | **0.61** |
| **6** | 0.89 | 0.83 | **0.61** | **0.72** |
| **7** | 1.00 | 0.89 | **0.67** | **0.78** |
| **8** | 0.94 | 0.94 | 0.94 | 0.89 |
| **9** | 0.94 | 1.00 | 0.83 | 0.83 |
| **10** | 0.83 | 0.83 | **0.61** | **0.67** |
| **11** | 0.94 | 1.00 | 0.89 | 0.94 |
| **12** | 1.00 | 0.94 | **0.78** | **0.78** |
| **13** | 0.89 | 0.89 | 0.89 | 0.89 |
| **14** | 0.94 | 0.94 | **0.72** | **0.72** |
| **15** | 0.94 | 0.94 | 0.89 | 0.94 |
| **16** | 1.00 | 1.00 | 0.94 | 0.89 |
| **17** | 0.89 | 0.83 | **0.72** | **0.78** |
| **18** | 0.94 | 0.94 | 0.94 | 0.89 |
| **19** | 0.94 | 0.94 | 0.89 | 0.89 |
| **20** | 0.94 | 0.94 | **0.78** | **0.78** |
| **21** | 1.00 | 1.00 | 0.83 | 0.83 |
| **22** | 1.00 | 0.89 | **0.78** | **0.78** |
| **23** | 0.89 | 0.89 | 0.83 | 0.83 |
| **24** | 0.89 | 0.89 | **0.72** | **0.78** |
| **25** | 1.00 | 0.94 | 0.83 | 0.89 |
| **26** | 1.00 | 1.00 | 0.94 | 0.94 |

| Ítems | Promedio ítem en Relevancia | Promedio ítem en Coherencia | Promedio ítem en Suficiencia | Promedio ítem en Claridad |
|---|---|---|---|---|
| **27** | 1.00 | 0.94 | **0.78** | **0.78** |
| **28** | 0.94 | 1.00 | 0.89 | 0.83 |
| **29** | 0.94 | 1.00 | 0.89 | 0.94 |
| **30** | 0.89 | **0.78** | **0.61** | **0.61** |
| **31** | 0.94 | 0.89 | **0.50** | **0.61** |
| **32** | 0.94 | 1.00 | 0.89 | 0.83 |
| **33** | 0.89 | 0.89 | **0.72** | **0.72** |
| **34** | 0.89 | 0.94 | **0.67** | **0.72** |

Nota: Elaboración propia con el software JAMOVI.

La lectura de la tabla revela un patrón definido. La variabilidad no se distribuye de manera uniforme entre los cuatro criterios, sino que se concentra en Claridad (V = 0.80) y Suficiencia (V = 0.81). Aunque los promedios globales de ambas dimensiones se sitúan dentro del umbral de aceptación, varios reactivos individuales quedan por debajo de él. Los ítems 1, 5, 10, 30 y 31 acumulan los valores más bajos, y el ítem 31 representa el caso más severo con 0.61 en Claridad y 0.50 en Suficiencia, lo que indica deficiencias estructurales en la formulación de la tarea que afectarían la comprensión del estudiante, comprometiendo su respuesta en este ítem.

Esta discrepancia entre los niveles de acuerdo en Relevancia y Coherencia, por un lado, y Claridad y Suficiencia por otro, adquiere mayor peso cuando se examina a través del coeficiente W de Kendall. El cálculo para el criterio de Claridad arrojó un valor cercano a cero (0.001), lo que revela que, más allá del promedio aceptable, los jueces no compartieron un criterio común para evaluarla. Mientras algunos penalizaron severamente la redacción de ciertos ítems, otros los valoraron positivamente. Este desacuerdo es un dato técnico relevante, pues sugiere que en el campo del diseño visual la claridad de una instrucción evaluativa tiene una dimensión contextual y subjetiva, difícil de capturar con un coeficiente estadístico.

### 4.2.1. Análisis cualitativo de las observaciones de los jueces evaluadores

Las observaciones abiertas registradas por los jueces en la matriz de validación (véase Anexo F) permitieron identificar oportunidades de mejora que el coeficiente V de Aiken no logra detectar por sí solo, dado que este indicador cuantifica el nivel de acuerdo, pero no explica sus causas. Del análisis de esas observaciones emergieron tres tipos recurrentes de señalamiento.

El primero concierne a la ambigüedad en el estímulo visual. En los ítems 24 y 29, los jueces indicaron que las imágenes de referencia contenían información irrelevante que podía desviar la atención del estudiante del objetivo evaluativo específico. El segundo tipo de señalamiento involucra la especificidad terminológica: en el ítem 6, relacionado con bases de datos de imágenes, dos expertos advirtieron que el uso de nombres propios de plataformas —como Biblioteca Digital Mundial— podría estar evaluando memoria antes que competencia de búsqueda, y sugirieron reorientar el ítem hacia criterios de selección genéricos como licencias, resolución y formato. El tercero atañe al encuadre de la tarea. Varios jueces recomendaron explicitar el rol del evaluado en el enunciado —mediante fórmulas del tipo "Usted es el diseñador encargado de..."— argumentando que la ausencia de ese contexto dificulta la toma de decisiones éticas y reduce la autenticidad de la situación evaluativa.

Los tres tipos de observación apuntan a un mismo problema de fondo. La diferencia entre los altos niveles de acuerdo en Relevancia y los bajos en Claridad no es aleatoria, sino que refleja una tensión propia de la evaluación en artes y diseño. Como señala Kędra (2018), la lectura visual implica decodificar signos e interpretar contextos culturales y estéticos de manera simultánea, de modo que formular una instrucción que oriente, sin restringir la interpretación, constituye un problema técnico genuino. Esta tensión se hace más evidente en los ítems del Estándar 3 de la ACRL, donde la interpretación de imágenes admite múltiples lecturas válidas, y se atenúa en los

reactivos de naturaleza normativa —como los asociados a ética visual y derechos de autor— donde el criterio de análisis es más unívoco. El hallazgo orienta con precisión el trabajo de revisión del instrumento hacia la reformulación de enunciados y estímulos visuales, de manera que la tarea guíe al estudiante hacia el criterio de evaluación específico sin reducir la complejidad interpretativa que justifica el ítem.

### 4.3. Análisis de Validez de la Estructura Interna

La validez de contenido establecida por los jueces es una condición necesaria pero no suficiente. Para determinar si los ítems funcionan de manera coherente cuando los responden estudiantes reales, se analizó la estructura interna del instrumento a partir de los datos del piloto. El propósito central fue verificar si el conjunto de reactivos opera de manera consistente y cohesiva para evaluar la alfabetización visual, y confirmar el supuesto de unidimensionalidad requerido para la posterior calibración bajo el Modelo de Rasch.

La muestra estuvo conformada por 54 estudiantes activos de noveno semestre del programa de Diseño Digital y Multimedia, 28 mujeres y 26 hombres, con edades comprendidas entre los 23 y 26 años. Los datos recolectados se procesaron con el software JAMOVI (versión 2.7.15), mediante los módulos especializados *snowIRT* y *Rasch Models*, calculando el Índice de Dificultad (propoción de aciertos, ***p***) y el Índice de Discriminación (correlación punto-biserial, ***rpbis***).

#### 4.3.1. Análisis de la Dificultad (p)

El índice de dificultad (p) expresa la proporción de estudiantes que respondieron correctamente cada ítem, con valores teóricos entre 0.00 —nadie acertó— y 1.00 —todos acertaron—. El análisis reveló una tendencia general hacia la facilidad de la prueba. El promedio de dificultad global se situó por encima de 0.70, lo que significa que la mayoría de los reactivos fueron resueltos correctamente por más del 70% de los estudiantes. Los ítems correspondientes a

tareas de reconocimiento básico o búsqueda de información alcanzaron valores cercanos a 0.90, comportamiento esperable en una muestra de noveno semestre, momento del programa en el que se asume que las competencias visuales fundamentales están consolidadas.

Desde la perspectiva psicométrica, este patrón configura un efecto techo parcial: el instrumento discrimina bien entre estudiantes de nivel básico e intermedio, pero pierde precisión en el extremo superior de la habilidad. Este fenómeno se examina con mayor detalle a través del Mapa de Wright en la sección 4.4.

### 4.3.2. Análisis de la Discriminación (rpbis)

La correlación punto-biserial (rpbis) evalúa la capacidad de cada ítem para distinguir a los estudiantes más competentes de los menos competentes. Un ítem con buena discriminación —comúnmente rpbis > 0.30— tiende a ser acertado por quienes obtienen puntajes altos en la prueba y fallado por quienes obtienen puntajes bajos. La prueba mostró índices positivos y aceptables en 32 de los 34 ítems, validando su consistencia interna general. Sin embargo, los ítems 1 y 20 presentaron coeficientes negativos, lo que constituye una anomalía que requiere atención específica.

Una discriminación negativa indica que los estudiantes con mayor puntaje global tienden a fallar el ítem más que los de bajo desempeño, lo que señala un error en la clave o una ambigüedad que confunde al evaluado más experto. El análisis cualitativo confirmó problemas de construcción en ambos casos: en el ítem 1, las diferencias entre opciones eran demasiado sutiles para aparecer como primera pregunta; en el ítem 20, la opción correcta era notablemente más corta que los distractores, lo que introdujo un sesgo de respuesta. Ante esa evidencia, ambos reactivos se excluyeron de los análisis subsiguientes, y la versión depurada quedó conformada por 32 ítems.

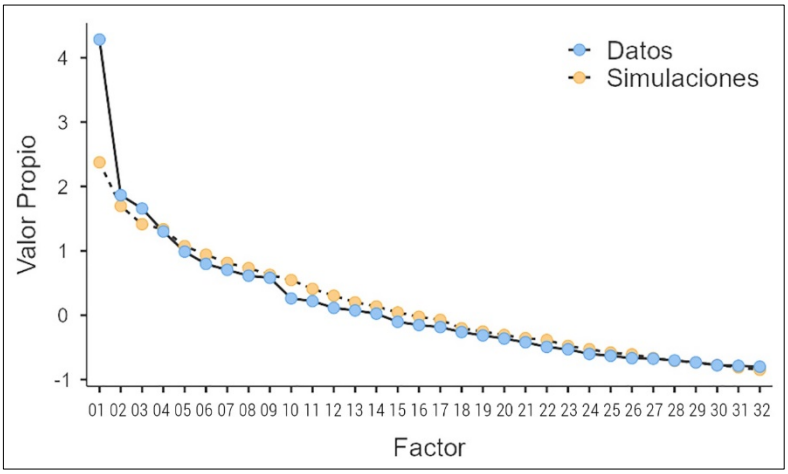
De manera adicional, los ítems 6, 14, 19 y 30 mostraron discriminaciones cercanas a cero, lo que indica que su aporte de información para diferenciar entre estudiantes es prácticamente nulo. Estos reactivos no comprometen la coherencia del instrumento como los negativos, pero reducen su eficiencia y requieren revisión.

### 4.3.3. Análisis de Dimensionalidad

Antes de calcular la consistencia interna de la prueba depurada, fue necesario verificar si el instrumento mide una variable latente predominante o si los ítems se agrupan en torno a factores independientes sin conexión entre sí. Para ello se ejecutó un análisis factorial exploratorio cuyos resultados se visualizan en el Gráfico de Sedimentación (*Scree Plot*) de la Figura 1.

**Figura 1.**

*Gráfico de sedimentación (Scree Plot) indicando la estructura dimensional de la prueba.*




Nota: elaboración propia con el software JAMOVI.

El gráfico muestra un cambio abrupto de pendiente —un "codo"— inmediatamente después del primer componente, cuyo valor propio supera con claridad al de todos los factores subsiguientes. Esta configuración es la señal característica de la unidimensionalidad esencial: existe una variable latente dominante que permea el desempeño en los tres bloques temáticos de búsqueda, lectura y escritura visual, y ética. La ligera inclinación descendente de la recta posterior

al codo, en lugar de un trazado perfectamente horizontal, es un fenómeno esperable en instrumentos de evaluación del desempeño ante constructos complejos. Sugiere la presencia de correlaciones residuales menores entre ítems —posiblemente atribuibles a formatos de tarea similares o micro-competencias específicas—, que carecen de la fuerza suficiente para constituir dimensiones independientes. Este hallazgo respalda tanto el uso de un puntaje global para interpretar los resultados como la viabilidad del Modelo de Rasch para la calibración definitiva.

### 4.3.4. Estimación de la Confiabilidad

Confirmada la unidimensionalidad esencial, se calculó la confiabilidad de la prueba depurada de 32 ítems mediante el coeficiente Kuder-Richardson 20 (KR-20), equivalente al Alfa de Cronbach para ítems dicotómicos. El análisis arrojó un coeficiente de 0.758, frente al valor inicial de 0.57 obtenido con los 34 ítems originales. La diferencia entre ambos valores ilustra el efecto concreto de la depuración: los ítems 1 y 20 introducían ruido suficiente como para reducir en veinte puntos la coherencia global del instrumento.

De acuerdo con los estándares de la *American Educational Research Association* et al. (2014) y los criterios de George y Mallery (2003), valores entre 0.70 y 0.80 son aceptables para fines de investigación y diagnóstico grupal. El coeficiente obtenido confirma que el 77% de la varianza en los puntajes observados corresponde a la habilidad real de los estudiantes en alfabetización visual, con el error de medición controlado dentro de márgenes razonables para una primera versión del prototipo.

La Tabla 3 sintetiza los estadísticos de dificultad y discriminación para los 34 ítems originales, junto con el diagnóstico técnico de cada reactivo. Un ítem se clasifica como funcional, muy bueno u óptimo cuando presenta un nivel de dificultad pertinente para la población evaluada y una discriminación positiva que demuestra su capacidad para diferenciar entre niveles de

competencia. Se clasifica como deficiente, débil o crítico cuando su aporte de información es nulo o cuando su discriminación contradice la lógica de la medición, independientemente de si resultó fácil o difícil para la muestra.

**Tabla 2**.

*Estadísticos descriptivos de los ítems: Dificultad y Discriminación.*

| Ítem | Dificultad (*p*) | Discriminación (*rpbis*) | Diagnóstico Técnico |
|---|---|---|---|
| **1** | **0.25** | **-0.17** | **Crítico.** Discriminación negativa; ítem defectuoso. |
| 2 | 0.81 | 0.32 | Funcional, nivel de dificultad bajo. |
| 3 | 0.91 | 0.31 | Muy fácil, discriminación aceptable. |
| 4 | 0.91 | 0.25 | Muy fácil, discriminación moderada. |
| 5 | 0.45 | 0.44 | **Óptimo.** Excelente equilibrio dificultad/discriminación. |
| **6** | 0.72 | **0.03** | **Deficiente.** Aporta información nula para diferenciar. |
| 7 | 0.89 | 0.14 | Fácil con baja capacidad discriminativa. |
| 8 | 0.83 | 0.46 | Muy buen discriminador. |
| 9 | 0.87 | 0.35 | Funcional. |
| 10 | 0.60 | 0.36 | Funcional, dificultad media. |
| **11** | **0.94** | 0.33 | **Techo.** Extremadamente fácil (51/54 aciertos). |
| 12 | 0.58 | 0.14 | Discriminación baja. |
| 13 | 0.77 | 0.27 | Funcional. |
| **14** | 0.43 | **0.01** | **Deficiente.** Discriminación nula pese a dificultad media. |

| Ítem | Dificultad (*p*) | Discriminación (*rpbis*) | Diagnóstico Técnico |
|---|---|---|---|
| 15 | 0.72 | 0.22 | Funcional. |
| 16 | 0.96 | 0.46 | Muy fácil, pero discrimina bien en niveles bajos. |
| 17 | 0.75 | 0.11 | Discriminación baja. |
| 18 | 0.74 | 0.22 | Funcional. |
| **19** | **0.96** | **0.05** | **Deficiente.** Ítem trivial sin poder discriminativo. |
| **20** | **0.23** | **-0.08** | **Crítico.** Discriminación negativa; posible error de clave. |
| 21 | 0.58 | 0.17 | Aceptable. |
| 22 | 0.72 | 0.19 | Aceptable. |
| 23 | 0.72 | 0.32 | Funcional. |
| 24 | 0.64 | 0.33 | Funcional. |
| 25 | 0.94 | 0.31 | Muy fácil. |
| 26 | 0.89 | 0.25 | Fácil. |
| 27 | 0.91 | 0.42 | Fácil con buena discriminación. |
| **28** | **0.75** | **0.54** | **Excelente.** Ítem con mayor poder discriminativo. |
| 29 | 0.85 | 0.50 | Muy bueno. |
| **30** | 0.83 | **0.07** | **Débil.** Requiere revisión urgente. |
| 31 | 0.68 | 0.33 | Funcional. |
| 32 | 0.87 | 0.36 | Funcional. |
| 33 | 0.79 | 0.47 | Muy bueno. |
| 34 | 0.68 | 0.45 | Muy bueno. |

Nota: Elaboración propia con el software JAMOVI.

## 4.4. Calibración Psicométrica bajo el Modelo de Rasch

Depurada la matriz de respuestas mediante los filtros clásicos de la TCT, los datos de los 32 ítems retenidos se sometieron a calibración bajo el Modelo de Rasch con dos propósitos: verificar que los ítems se ajustan al modelo probabilístico y analizar si la dificultad del instrumento corresponde al nivel de competencia de la población evaluada.

A diferencia de la TCT, que trabaja con puntuaciones ordinales derivadas de la sumatoria simple de aciertos, el Modelo de Rasch transforma los puntajes brutos en medidas expresadas en logits, una unidad que permite ubicar la dificultad de los ítems y la habilidad de los estudiantes en una misma escala continua. Esta propiedad hace posible determinar, con independencia de las características particulares de la muestra, si un reactivo específico está al alcance, por encima o por debajo de la competencia de un estudiante dado (Bond & Fox, 2015). Los resultados se analizan en dos momentos: el ajuste de los ítems al modelo y la distribución conjunta persona-ítem visualizada en el Mapa de Wright.

### 4.4.1. Ajuste de los ítems al modelo (*Infit* y *Outfit*)

El Modelo de Rasch establece que un estudiante con alta habilidad tendrá mayor probabilidad de acertar los ítems fáciles, mientras que uno con baja habilidad tenderá a fallar los reactivos complejos. Las desviaciones frente a ese patrón esperado se cuantifican mediante dos índices de ajuste. El *Infit* o ajuste interno, que evalúa si el estudiante responde coherentemente a las preguntas situadas justo en su nivel de competencia, es sensible a inconsistencias en la zona central de la escala donde se concentra la mayor parte de la información. El *Outfit* o ajuste externo, en cambio, detecta eventos improbables en los extremos: el acierto por azar de un estudiante de bajo nivel en una pregunta muy difícil, o el fallo inexplicable de un estudiante experto ante una pregunta muy fácil. Ambos índices se expresan como estadísticos de ajuste cuyo rango productivo

para pruebas de investigación oscila entre 0.5 y 1.5, según los parámetros establecidos por Linacre (1994). Valores superiores a 1.5 indican aleatoriedad excesiva, mientras que valores inferiores a 0.5 sugieren una redundancia o sobre dependencia tan determinista entre ítems que el modelo pierde su capacidad de discriminación. Es decir que, si un ítem tiene un índice muy bajo, (inferior a 0.5) significa que los estudiantes con alta habilidad siempre lo aciertan y los de baja habilidad siempre lo fallan de forma casi mecánica. La Tabla 4 reporta las medidas de dificultad en logits, el error estándar de estimación asociado a cada medida —donde valores más bajos indican mayor precisión en la calibración— y los índices de ajuste *Infit* y *Outfit* para los 32 ítems calibrados.

**Tabla 3.**

*Estadísticos de ajuste al Modelo Rasch: Medida, Infit y Outfit.*

| Ítem | Measure Medida | S.E.Measure Error Estándar de Medida | Infit MNSQ Ajuste Interno | Outfit MNSQ Ajuste Externo |
|---|---|---|---|---|
| **I02** | -1.668 | 0.37 | 0.96 | 1.1 |
| **I03** | -2.551 | 0.49 | 0.929 | 0.973 |
| **I04** | -2.551 | 0.49 | 0.975 | 1.081 |
| **I05** | 0.268 | 0.292 | 0.882 | 0.881 |
| **I06** | -1.076 | 0.323 | 1.194 | 1.384 |
| **I07** | -2.33 | 0.453 | 1.04 | 1.152 |
| **I08** | -1.81 | 0.385 | 0.906 | 0.773 |
| **I09** | -1.966 | 0.403 | 0.953 | 0.915 |
| **I10** | -0.415 | 0.296 | 0.957 | 0.925 |
| **I11** | -3.145 | 0.614 | 0.897 | 0.781 |
| **I12** | -0.328 | 0.294 | 1.089 | 1.07 |
| **I13** | -1.294 | 0.338 | 1.014 | 1.016 |
| **I14** | 0.353 | 0.293 | 1.189 | 1.227 |
| **I15** | -1.076 | 0.323 | 1.07 | 1.021 |
| **I16** | -3.595 | 0.739 | 0.833 | 0.44 |
| **I17** | -1.183 | 0.33 | 1.098 | 1.288 |
| **I18** | -1.183 | 0.33 | 1.073 | 1.051 |
| **I19** | -3.595 | 0.739 | 1.054 | 1.496 |
| **I21** | -0.328 | 0.294 | 1.085 | 1.21 |
| **I22** | -1.076 | 0.323 | 1.067 | 1.123 |
| **I23** | -1.076 | 0.323 | 0.998 | 0.972 |

| | | | | |
|---|---|---|---|---|
| **I24** | -0.684 | 0.304 | 0.988 | 1.008 |
| **I25** | -3.145 | 0.614 | 0.906 | 0.931 |
| **I26** | -2.33 | 0.453 | 1.025 | 0.858 |
| **I27** | -2.551 | 0.49 | 0.884 | 0.728 |
| **I28** | -1.294 | 0.338 | 0.848 | 0.776 |
| **I29** | -1.966 | 0.403 | 0.848 | 0.788 |
| **I30** | -1.668 | 0.37 | 1.139 | 1.288 |
| **I31** | -0.874 | 0.312 | 1.008 | 0.99 |
| **I32** | -2.137 | 0.425 | 0.931 | 0.874 |
| **I33** | -1.535 | 0.358 | 0.879 | 0.863 |
| **I34** | -0.778 | 0.308 | 0.906 | 0.867 |

Nota: Elaboración propia con el software JAMOVI.

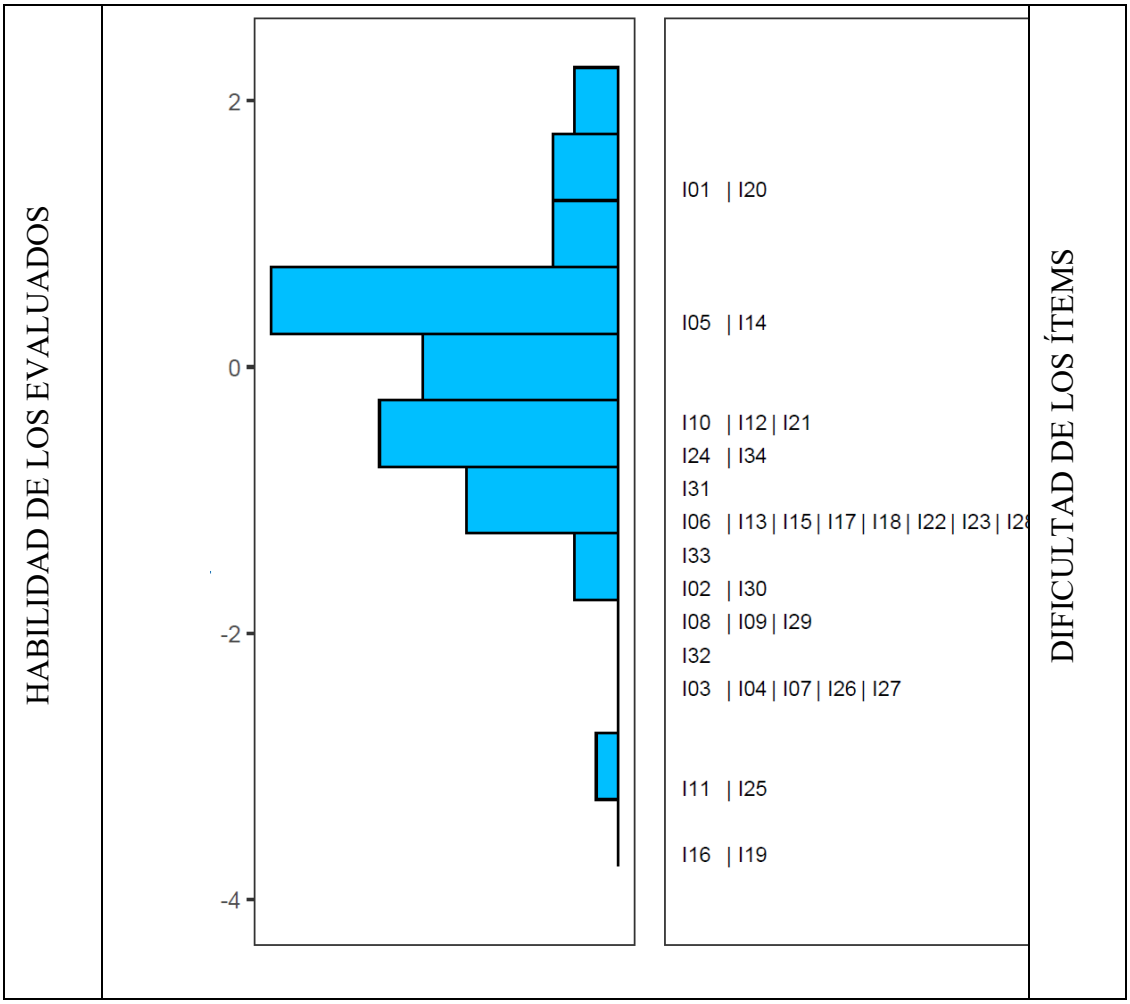
El análisis de los índices reportados en la Tabla 4 muestra que la totalidad de los reactivos se mantuvo dentro del rango productivo de medida, sin que se detectaran ítems que generaran ruido excesivo ni comportamientos erráticos en las respuestas de los estudiantes. Este comportamiento estadístico confirma la validez de constructo desde una perspectiva probabilística: los 32 reactivos contribuyen de manera consistente a la medición de la alfabetización visual y operan con la coherencia que el modelo exige.

### 4.4.2. Mapa de Wright: distribución ítem-persona y efecto techo

La distribución conjunta de habilidades e ítems se visualiza en el Mapa de Wright (Figura 2), que sitúa simultáneamente las frecuencias de habilidad de los estudiantes —eje izquierdo— y la dificultad calibrada de los ítems —eje derecho— a lo largo del continuo logit. Los valores positivos superiores indican mayor competencia o mayor exigencia técnica; los valores negativos inferiores corresponden a niveles elementales.

**Figura 2.**

*Mapa de Wright: Distribución conjunta de habilidad de los estudiantes y dificultad de los ítems.*




Nota: Elaboración propia con el software JAMOVI.

Por convención metodológica, el Modelo de Rasch fija el promedio de dificultad de los ítems en 0.0 logits, valor que actúa como punto de anclaje de la escala. Al contrastar ese centro con la ubicación de los participantes, se identificó un desfase positivo: mientras la exigencia de la prueba promedió 0.0 logits, la habilidad promedio de la muestra alcanzó los 0.90 logits. En términos gráficos, la distribución de los estudiantes aparece desplazada hacia la zona superior

respecto a la columna de reactivos, lo que confirma que el nivel de competencia promedio del grupo supera la exigencia promedio del instrumento.

Este fenómeno, tipificado como efecto techo parcial, tiene dos lecturas que conviene distinguir. Desde la psicometría, es una limitación técnica: el instrumento discrimina con precisión en los niveles básico e intermedio, pero pierde resolución en el extremo superior de la habilidad, donde la escasez de reactivos calibrados por encima de +2.0 logits impide distinguir diferencias sutiles entre los estudiantes de desempeño sobresaliente. Esta limitación implica que la siguiente versión del instrumento deberá incorporar reactivos de alta complejidad hermenéutica anclados en situaciones visuales que el currículo no aborda explícitamente. Desde la evaluación educativa, el mismo dato es positivo pues establece que los estudiantes de noveno semestre dominan con holgura los ítems técnicos y éticos del instrumento indicando que el programa ha consolidado esas competencias de forma efectiva en su perfil de egreso. La zona de mayor variabilidad en los resultados, como muestra el mapa, corresponde a los reactivos del bloque de lectura visual crítica, hallazgo cuyas implicaciones pedagógicas se desarrollan en la sección 4.6.

### 4.5. Contraste entre Dificultad Teórica y Dificultad Empírica

Durante el diseño del instrumento, los ítems se clasificaron por nivel de complejidad cognitiva según la Taxonomía de Bloom. La calibración Rasch permite ahora confrontar esa clasificación teórica con la dificultad que los reactivos representaron efectivamente para los estudiantes. El procedimiento consistió en categorizar las estimaciones Rasch en tres niveles empíricos —bajo para valores menores a -0.5 logits, medio para el rango entre -0.5 y +0.5 logits, y alto para valores superiores a +0.5 logits— y contrastarlos con la matriz de especificaciones original. La Tabla 5 presenta el detalle de esta comparación.

**Tabla 4.**

*Matriz de comparación: Nivel Bloom vs. Nivel Rasch y correspondencia.*

| Ítem | Medida Rasch (Logits) | Nivel Rasch Empírico Detectado | Nivel Bloom (Teórico - Diseño) | Correspondencia |
|---|---|---|---|---|
| **I05** | 0.268 | MEDIO (El más difícil) | MEDIO | No (Más fácil de lo esperado) |
| **I14** | 0.353 | MEDIO (El más difícil) | ALTO | No (Más fácil de lo esperado) |
| **I10** | -0.415 | MEDIO | ALTO | No (Más fácil de lo esperado) |
| **I12** | -0.328 | MEDIO | BAJO | No (Más Difícil de lo esperado) |
| **I21** | -0.328 | MEDIO | ALTO | No (Más Difícil de lo esperado) |
| **I24** | -0.684 | BAJO | MEDIO | No (Más fácil de lo esperado) |
| **I34** | -0.778 | BAJO | BAJO | Sí |
| **I31** | -0.874 | BAJO | ALTO | No (Más fácil de lo esperado) |
| **I06** | -1.076 | BAJO | ALTO | No (Más fácil de lo esperado) |
| **I15** | -1.076 | BAJO | BAJO | No (Más fácil de lo esperado) |
| **I22** | -1.076 | BAJO | MEDIO | No (Más fácil de lo esperado) |
| **I23** | -1.076 | BAJO | BAJO | Sí |
| **I17** | -1.183 | BAJO | BAJO | Sí |
| **I18** | -1.183 | BAJO | BAJO | Sí |
| **I13** | -1.294 | BAJO | ALTO | No (Más fácil de lo esperado) |
| **I28** | -1.294 | BAJO | BAJO | Sí |
| **I33** | -1.535 | BAJO | MEDIO | No (Más fácil de lo esperado) |
| **I02** | -1.668 | BAJO | ALTO | No (Más fácil de lo esperado) |
| **I30** | -1.668 | BAJO | MEDIO | No (Más fácil de lo esperado) |
| **I08** | -1.81 | BAJO | BAJO | Sí |

| Ítem | Medida Rasch (Logits) | Nivel Rasch Empírico Detectado | Nivel Bloom (Teórico - Diseño) | Correspondencia |
|---|---|---|---|---|
| **I09** | -1.966 | BAJO | MEDIO | No (Más fácil de lo esperado) |
| **I29** | -1.966 | BAJO | ALTO | No (Más fácil de lo esperado) |
| **I32** | -2.137 | BAJO | BAJO | Sí |
| **I07** | -2.33 | BAJO | MEDIO | No (Más fácil de lo esperado) |
| **I26** | -2.33 | BAJO | BAJO | Sí |
| **I03** | -2.551 | BAJO | MEDIO | No (Más fácil de lo esperado) |
| **I04** | -2.551 | BAJO | BAJO | Sí |
| **I27** | -2.551 | BAJO | ALTO | No (Más fácil de lo esperado) |
| **I11** | -3.145 | BAJO | BAJO | Sí |
| **I25** | -3.145 | BAJO | BAJO | Sí |
| **I16** | -3.595 | BAJO | MEDIO | No (Más fácil de lo esperado) |
| **I19** | -3.595 | BAJO | MEDIO | No (Más fácil de lo esperado) |

Nota: Elaboración propia con el software JAMOVI.

La coincidencia directa entre el nivel planeado y el observado se situó en un 40.6%, concentrada casi en su totalidad en el nivel bajo, donde el diseño predijo correctamente la dificultad en 12 de los 13 casos que coincidieron. El dato más llamativo es que ningún reactivo alcanzó la categoría de dificultad alta bajo los parámetros de Rasch, a pesar de que el diseño original contemplaba un 28% de ítems en ese nivel, orientados a procesos de evaluación y creación. La totalidad de las tareas concebidas como de alta complejidad funcionó, en la práctica, como reactivos de dificultad media o baja para esta muestra.

Una primera lectura de esta discrepancia podría interpretarse como un error en el diseño del instrumento. Sin embargo, los datos apuntan a una conclusión más interesante sobre el

programa. La Taxonomía de Bloom clasifica la complejidad del proceso mental que el ítem pretende activar, pero no controla el grado de familiaridad que los estudiantes tienen con el contenido evaluado. Un ítem diseñado para medir análisis crítico puede resultar empíricamente fácil cuando el criterio de análisis que exige forma parte del vocabulario cotidiano del programa. Lo que la comparación revela, entonces, no es que los ítems estén mal diseñados, sino que los estudiantes de noveno semestre del programa de Diseño Digital y Multimedia han interiorizado como competencias de ejercicio rutinario procesos que la taxonomía clasifica como cognitivamente complejos. Dicho de otra manera, el programa ha sido lo suficientemente efectivo como para que sus graduandos operen con naturalidad en niveles de pensamiento que en otros contextos se considerarían exigentes.

Este hallazgo tiene una implicación concreta para el desarrollo del instrumento. Si en versiones futuras se quiere mantener la capacidad discriminativa en los niveles altos de habilidad, la dificultad de los ítems de mayor complejidad no debería asociarse únicamente con el nivel taxonómico del proceso cognitivo requerido, sino con situaciones visuales inéditas o con problemas que el currículo no aborde explícitamente. La dificultad técnica necesaria para discriminar a los estudiantes sobresalientes en este programa deberá construirse sobre contenidos que estén deliberadamente fuera de su zona de familiaridad curricular.

### 4.6. La Triangulación como Mecanismo de Validación Convergente

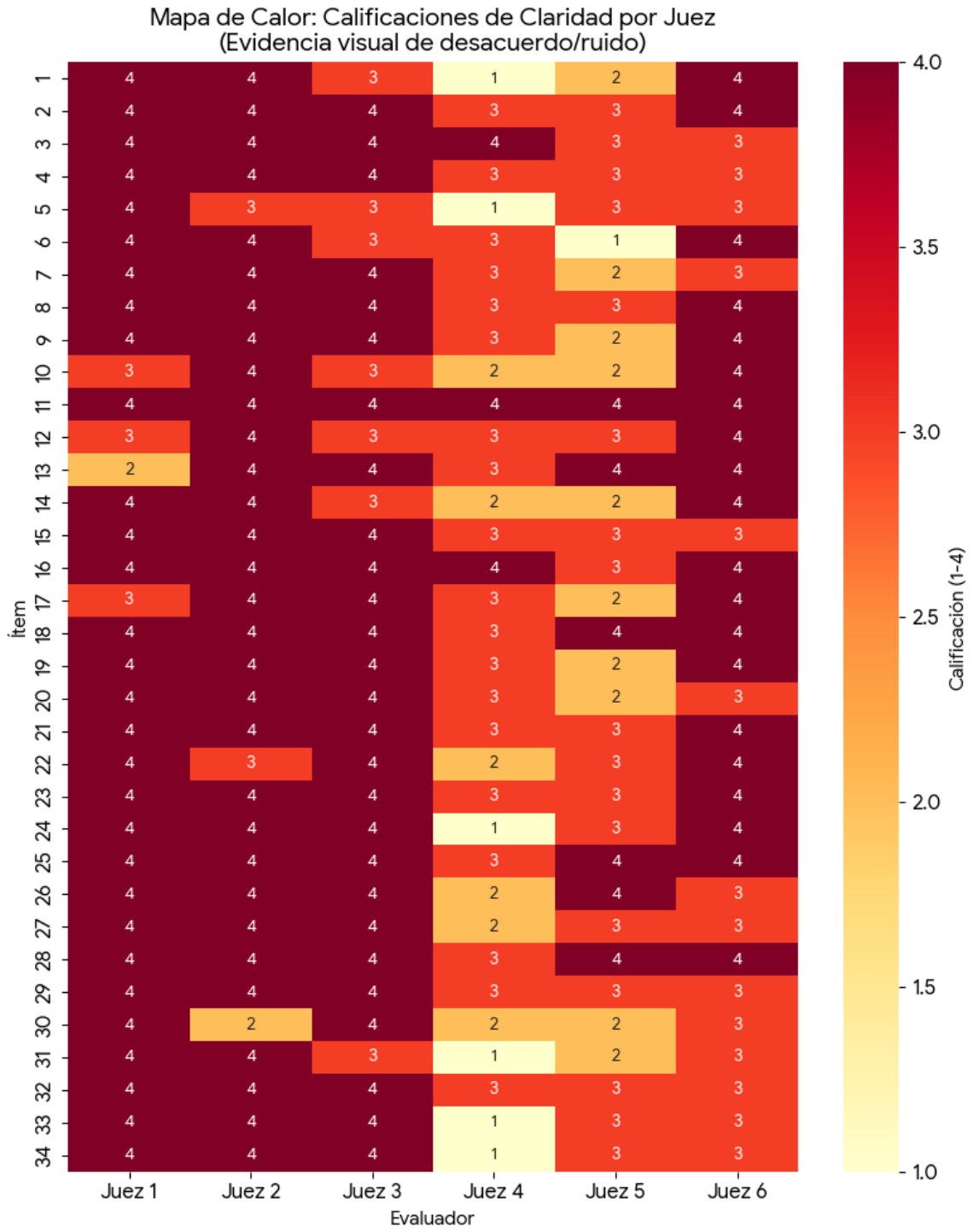
Con los resultados del juicio experto y del piloto presentados en las secciones anteriores, corresponde ahora contrastarlos directamente para identificar qué revela su coincidencia y qué revelan sus diferencias sobre la calidad del instrumento. Al ejecutar ambos procesos en paralelo, fue posible confrontar la validez teórica percibida por los expertos con la validez empírica demostrada por los estudiantes, en lugar de asumir que una implica la otra.

La convergencia más sólida entre ambas fuentes se produjo en la dimensión de Relevancia. El consenso de los jueces ($V > 0.90$) respecto a la pertinencia de los reactivos encontró respaldo en el comportamiento estadístico del piloto: los ítems que los expertos consideraron representativos del constructo funcionaron también con coherencia psicométrica, confirmando que la operacionalización de los estándares de la ACRL mediante el DCE fue exitosa. No hubo discrepancias sobre qué se debía evaluar. Las competencias de interpretación, evaluación crítica y uso ético de imágenes fueron reconocidas unánimemente como centrales para el perfil del diseñador digital, y los datos empíricos respaldaron esa lectura. Esta convergencia valida la Matriz de Especificaciones construida bajo el DCE y demuestra que es posible traducir los estándares de la ACRL en tareas evaluativas concretas con un grado de representatividad técnica defendible.

Las divergencias más instructivas se concentraron en la dimensión de Claridad. El análisis de concordancia mediante el coeficiente W de Kendall arrojó un valor cercano a cero (0.001), evidenciando que los jueces no compartieron un criterio común para evaluar esta dimensión: mientras algunos penalizaron severamente la redacción de ciertos reactivos, otros los valoraron de manera positiva. La Figura 3, que ilustra esa distribución mediante un mapa de calor, hace visible lo que el promedio global ocultaba.

**Figura 3.**

*Mapa de Calor (Calificaciones de Claridad por Juez) Coeficiente de Kendall.*




Nota: Elaboración propia con el software JAMOVI.

Lo más revelador no es el desacuerdo en sí mismo, sino lo que ocurrió al contrastarlo con el comportamiento empírico de esos mismos ítems. En algunos casos, la alerta de los jueces predijo con precisión el fallo psicométrico: el ítem 1, penalizado por varios expertos en Claridad, discriminó negativamente en el piloto. En otros casos, en cambio, los estudiantes lograron interactuar exitosamente con ítems que los expertos consideraron confusos. Este hallazgo respalda la visión de Kane (2013) sobre la validez como argumento interpretativo: el juicio experto formula una hipótesis que los datos deben confirmar o refutar. La triangulación permitió identificar que la claridad en la evaluación de competencias visuales no es una propiedad absoluta del enunciado, sino un atributo contextual que depende de las diferencias entre el marco interpretativo del experto y el del estudiante. Como señala Kędra (2018), la lectura visual implica decodificar signos e interpretar contextos culturales y estéticos de manera simultánea, de modo que un estímulo que un experto percibe como técnicamente impreciso puede ser una instrucción suficiente para quien opera cotidianamente con ese tipo de imágenes.

Esta tensión se hizo más evidente en los ítems del Estándar 3 de la ACRL, donde la interpretación de imágenes admite múltiples lecturas válidas, y se atenuó en los reactivos de naturaleza normativa asociados a ética visual y derechos de autor, donde el criterio de análisis es más unívoco. Los ítems técnicos y éticos mostraron los comportamientos psicométricos más estables —discriminación positiva y dificultad media— y obtuvieron mayor consenso entre los jueces. Los ítems interpretativos, en cambio, concentraron tanto el desacuerdo inter-jueces como la mayor variabilidad en el desempeño de los estudiantes. La Figura 4 sintetiza visualmente esta distribución al contrastar la valoración de claridad por parte de los expertos con la discriminación empírica de cada reactivo.

**Figura 4.**

*Gráfica Integrada de Triangulación: Piloto vs. Expertos.*


Nota: Elaboración propia con el software JAMOVI.

La gráfica confirma la hipótesis de triangulación. Los ítems ubicados en la zona crítica —abajo a la izquierda, como el 1 y el 20— fallaron tanto en la valoración teórica de claridad como en la discriminación empírica. La concentración de reactivos en la zona superior derecha valida la solidez del resto del instrumento. Más importante aún, la distribución revela que la objetividad en la evaluación de la alfabetización visual se alcanza cuando el enunciado precisa el criterio de análisis con suficiente claridad, para orientar la respuesta del estudiante sin reducir la riqueza interpretativa de la imagen. Ese equilibrio entre precisión enunciativa y complejidad evaluativa es,

justamente, lo que la triangulación entre juicio experto y evidencia empírica permite calibrar con un grado de certeza que ninguna de las dos fuentes alcanzaría por separado.

El efecto techo adquiere una lectura más precisa al triangularse con el juicio experto. Como limitación técnica, ya se estableció que el instrumento pierde resolución en el extremo superior de la habilidad; lo que la triangulación permite determinar es sobre qué dimensiones específicas esa pérdida es más relevante y, por consiguiente, dónde debe expandirse el banco de ítems con mayor prioridad. Los reactivos que los estudiantes resolvieron con mayor facilidad —los asociados a formatos de archivo, criterios de licencia e identificación de manipulación de imágenes— son precisamente los que los jueces calificaron con mayor consenso en Relevancia y Coherencia. El alto porcentaje de aciertos en esas tareas indica que el programa ha consolidado ese conocimiento con eficacia; agregar reactivos más exigentes en las dimensiones técnica y ética tiene, por tanto, menor urgencia que en la dimensión hermenéutica. Los ítems de mayor dificultad relativa pertenecen al bloque de lectura visual crítica: interpretación de significados, evaluación de la eficacia comunicativa de una imagen, análisis del contexto cultural de producción. Que esas tareas sean las que mayor variación producen en los puntajes, y que los expertos hayan mostrado mayor desacuerdo al evaluarlas, revela que la formación hermenéutica es menos homogénea que la formación técnica. Esta asimetría orienta con precisión la dirección en que el instrumento debe desarrollarse: no en todas las dimensiones por igual, sino prioritariamente en aquella que mayor variabilidad diagnóstica presenta.

### 4.7. Conclusión del Capítulo

Los resultados presentados en este capítulo permiten sostener que el instrumento ha superado la etapa exploratoria y cuenta con propiedades psicométricas suficientes para justificar su desarrollo en una fase de mayor escala. La estrategia de validación convergente demostró su

pertinencia metodológica: al ejecutar simultáneamente el juicio de expertos y la aplicación piloto, fue posible obtener un diagnóstico que ninguna de las dos fuentes habría producido por separado. La estructura de contenido del instrumento resultó sólida —con coeficientes V de Aiken superiores a 0.93 en Relevancia y Coherencia— y su comportamiento psicométrico confirmó una consistencia interna aceptable para una primera versión del prototipo (KR-20 = 0.758), junto con el ajuste satisfactorio de los 32 ítems depurados al modelo probabilístico de Rasch.

Las limitaciones identificadas a lo largo del análisis tienen un valor que va más allá de su función correctiva. La discriminación negativa de los ítems 1 y 20 orientó con precisión la revisión de los distractores. El efecto techo reveló que la formación técnica y ética del programa está consolidada, y señaló la lectura visual crítica como la dimensión que mayor atención curricular requiere. La muestra reducida, por su parte, es una condición propia de la fase exploratoria que el diseño adoptado supo aprovechar: los estadísticos de la TCT operaron como filtro diagnóstico eficiente precisamente porque la calibración Rasch definitiva se reserva para una etapa con mayor volumen de datos. En conjunto, estas limitaciones lejos de debilitar las conclusiones del estudio; establecen con exactitud el trabajo que resta para consolidar el instrumento.

El capítulo siguiente recoge las conclusiones generales de la investigación, discute sus implicaciones para la evaluación de la alfabetización visual en la educación superior en diseño y propone los lineamientos técnicos para la siguiente fase de desarrollo del instrumento.

# CAPÍTULO 5. CONCLUSIONES Y RECOMENDACIONES

Esta investigación respondió la pregunta formulada en el Capítulo 1 mediante el recorrido metodológico documentado en los capítulos anteriores. Este capítulo recoge las conclusiones que la evidencia sostiene, examina los hallazgos de mayor alcance para el campo y la institución, y traza la ruta de desarrollo del instrumento.

## 5.1. Conclusión General

El prototipo desarrollado en esta investigación demuestra que la alfabetización visual en estudiantes de diseño de último semestre puede evaluarse mediante un instrumento objetivo con fundamento psicométrico sólido. La validez de contenido fue confirmada por el panel de expertos, con coeficientes V de Aiken superiores a 0.90 en relevancia y coherencia. La consistencia interna de la versión depurada resultó aceptable para una primera versión del prototipo (KR-20 = 0.758), y la calibración bajo el Modelo de Rasch mostró que los 32 ítems retenidos se ajustan al modelo probabilístico sin generar ruido estadístico significativo. Estos tres resultados, obtenidos de fuentes distintas y contrastados mediante triangulación metodológica, convergen en una misma dirección: el instrumento cumple las condiciones técnicas para avanzar hacia una calibración definitiva a mayor escala.

Un resultado de particular alcance teórico es lo que el funcionamiento del instrumento revela sobre el constructo mismo. La obtención de un coeficiente de confiabilidad aceptable en un instrumento que evalúa simultáneamente competencias tan diversas como la búsqueda de imágenes, su interpretación crítica y su uso ético aporta evidencia empírica de que estas habilidades operan simultáneamente en la práctica cognitiva de los estudiantes de diseño. El estudiante competente en una dimensión tiende a serlo en las otras, lo que valida la posibilidad de construir una escala única —en logits— que ordene a los estudiantes desde niveles elementales

hasta niveles avanzados de alfabetización visual, sin necesidad de fragmentar la prueba en sub-escalas independientes.

## 5.2. Hallazgos con significado transversal

### 5.2.1. La alfabetización visual admite medición psicométrica rigurosa

El hallazgo que atraviesa toda la investigación es también el más difícil de dar por sentado al inicio de un estudio como este: que un constructo tan complejo y multidimensional como la alfabetización visual puede traducirse en un instrumento objetivo con propiedades psicométricas defendibles. La integración de los Estándares de Competencia en Alfabetización Visual de la ACRL (2011) con el modelo teórico de Kędra (2018) proporcionó el marco necesario para estructurar los siete estándares de la ACRL en un constructo con tres dimensiones evaluables mediante ítems de selección múltiple: la lectura visual, la escritura visual y la ética visual. La metodología del Diseño Centrado en Evidencias permitió traducir esas dimensiones en tareas concretas con una alineación constructiva verificable entre los estándares, los indicadores de desempeño y los reactivos del instrumento.

La viabilidad de esa traducción conceptual en reactivos concretos no era evidente a priori. La literatura especializada advierte sobre el riesgo de que instrumentos de este tipo se fragmenten en sub-escalas desconectadas, lo que haría imposible construir una medida global del constructo (Li y Potter, 2023). El coeficiente KR-20 de 0.758 obtenido en la versión depurada de 32 ítems aporta evidencia empírica de que ese riesgo no se materializó: las tres dimensiones evaluadas comparten una variable latente dominante que justifica su tratamiento como un constructo unitario. Este resultado es el que habilita técnicamente la calibración definitiva bajo el Modelo de Rasch, cuyo supuesto fundamental es precisamente la existencia de ese rasgo latente unificado.

### 5.2.2. La triangulación produjo un diagnóstico que ninguna fuente por separado habría generado

La decisión metodológica de ejecutar simultáneamente el juicio de expertos y la aplicación piloto, en lugar de aplicarlos de forma secuencial, más que una elección de diseño metodológico, fue la condición que hizo posible el diagnóstico más fino del instrumento. Al contrastar lo que los expertos consideraron que cada ítem medía con lo que el comportamiento estadístico de los estudiantes reveló que ese ítem efectivamente producía, emergieron convergencias y divergencias que ninguna de las dos fuentes habría podido identificar por separado.

La convergencia más sólida se produjo en la dimensión de relevancia. El consenso de los jueces respecto a la pertinencia de los reactivos encontró respaldo en el comportamiento empírico del piloto, confirmando que los estándares de la ACRL se convirtieron exitosamente en tareas evaluativas representativas del constructo. Las divergencias más evidentes se concentraron en la dimensión de claridad, donde el coeficiente W de Kendall cercano a cero (0.001) reveló que los jueces no compartían un criterio común. Al contrastar ese desacuerdo con el desempeño real de los estudiantes, fue posible distinguir los ítems cuya ambigüedad era genuinamente problemática —como los ítems 1 y 20, que discriminaron negativamente— de aquellos que los expertos consideraron confusos pero que los estudiantes resolvieron sin dificultad. Sin esa comparación directa, la decisión sobre cuáles reactivos eliminar y cuáles conservar habría dependido únicamente del criterio experto, con el riesgo de descartar ítems funcionalmente válidos.

Este hallazgo tiene implicaciones metodológicas que trascienden este estudio particular. En el campo de la evaluación de competencias visuales, donde la subjetividad del juicio experto es especialmente alta, la triangulación entre valoración teórica y evidencia empírica no es un refinamiento opcional sino una condición de rigor. Como sostiene Kane (2013), la validez es un

argumento que se construye acumulando evidencias desde distintos ángulos; ninguna fuente individual, por autorizada que sea, puede sostenerlo por sí sola.

### 5.2.3. El instrumento reveló una asimetría curricular con implicaciones pedagógicas concretas

El instrumento demostró su utilidad diagnóstica más allá de sus propiedades psicométricas: al confrontar los patrones de respuesta con los estándares que cada ítem busca medir, reveló una asimetría en el perfil de competencias de los estudiantes evaluados que las evaluaciones curriculares ordinarias no habían identificado. Ese hallazgo es consecuencia directa del efecto techo documentado en el Capítulo 4. Como limitación técnica, el efecto techo indica que el instrumento pierde resolución en el extremo superior de la habilidad; como dato diagnóstico, lo que esa pérdida de resolución revela —al examinar en qué dimensiones se concentra— es una distribución desigual de competencias en el perfil de egreso del programa.

Los reactivos que los estudiantes de noveno semestre resolvieron con mayor facilidad corresponden a las dimensiones técnica y ética del constructo: formatos de archivo, criterios de licencia, identificación de manipulación de imágenes. El alto porcentaje de aciertos en esos ítems indica que el programa ha consolidado ese conocimiento de manera efectiva; la pérdida de resolución del instrumento en esa zona no representa, por tanto, la prioridad de desarrollo del banco de ítems.

Los reactivos de mayor dificultad relativa pertenecen al bloque de lectura visual crítica: interpretación de significados, evaluación de la eficacia comunicativa de una imagen, análisis del contexto cultural de producción. Que precisamente esas tareas sean las que mayor variación producen en los puntajes, y que los expertos hayan mostrado mayor desacuerdo al evaluarlas, indica que la formación hermenéutica es menos homogénea que la formación técnica. La lectura

pedagógica indica que mientras todos los estudiantes evaluados conocen los formatos de archivo y los criterios de licencia, no todos comparten los mismos marcos interpretativos para juzgar la eficacia comunicativa de una imagen o para analizar su contexto cultural de producción.

El contraste entre la dificultad teórica prevista según la Taxonomía de Bloom y la empíricamente medida por el Modelo de Rasch precisa esta lectura. Ningún reactivo alcanzó la categoría de dificultad alta bajo los parámetros del modelo de Rasch, a pesar de que el diseño original contemplaba un 28% de ítems en ese nivel. Esto no indica un error de diseño sino una característica del programa: los estudiantes de noveno semestre han interiorizado como competencias de ejercicio rutinario procesos que la taxonomía clasifica como cognitivamente complejos. El instrumento, al hacer visible la brecha entre la complejidad taxonómica y la familiaridad curricular, aporta información que ningún mecanismo de evaluación disponible ordinariamente en el programa podría producir con este nivel de precisión. Esa es su contribución diagnóstica más concreta.

Esta asimetría tiene implicaciones pedagógicas que trascienden la calibración del instrumento. Que las competencias técnicas y éticas estén consolidadas mientras la lectura visual crítica presenta mayor dispersión sugiere una diferencia en la manera como el currículo aborda cada dimensión. Las competencias técnicas —formatos de archivo, resolución, criterios de licencia— se enseñan mediante contenidos declarativos con respuestas verificables, lo que facilita tanto su instrucción como su evaluación. La ética visual, aunque involucra juicios de valor, se apoya en marcos normativos —derechos de autor, licencias Creative Commons— que proporcionan criterios de análisis relativamente estables. La lectura visual crítica, en cambio, exige operaciones cognitivas de naturaleza distinta: interpretar la intencionalidad de una composición, evaluar la eficacia comunicativa de una imagen en función de su audiencia, o analizar cómo el

contexto cultural de producción condiciona el significado. Estas son competencias que se adquieren por exposición sostenida a ejercicios de interpretación con retroalimentación experta, y que el currículo posiblemente desarrolla de forma más implícita que sistemática.

La implicación curricular más directa es que el programa podría beneficiarse de espacios pedagógicos dedicados específicamente al análisis hermenéutico de imágenes como electivas de profundización sobre lectura visual donde los estudiantes confronten interpretaciones divergentes sobre un mismo estímulo, integrando en cursos específicos, ejercicios de evaluación crítica de campañas visuales reales, o análisis comparativos de cómo diferentes contextos culturales producen y consumen las mismas imágenes. El instrumento desarrollado en esta investigación ofrece un recurso concreto para orientar esas decisiones: al aplicarse en dos momentos de la carrera —por ejemplo, al inicio del ciclo de profundización en sexto semestre y al final del programa—, permitiría medir el valor agregado del currículo en cada dimensión de la alfabetización visual y verificar si los ajustes pedagógicos introducidos producen el efecto esperado. Esta necesidad se ha intensificado con la consolidación de la IA generativa en los entornos visuales cotidianos. Hill y Conceição (2025) advierten que la proliferación de deepfakes y de medios manipulados algorítmicamente desafía la confianza pública y exige fortalecer la alfabetización crítica visual en la educación superior. En ese escenario, formar diseñadores capaces de leer críticamente imágenes es una contribución directa a la calidad del ecosistema informativo y la cultura visual del país.

### 5.3. Alcance y condiciones del estudio

Los resultados de esta investigación tienen un alcance definido por las condiciones metodológicas en que se produjeron. Tres de esas condiciones restringen la generalización de las conclusiones de manera específica y deben explicitarse antes de proyectar cualquier aplicación del instrumento a contextos o poblaciones distintos.

La primera concierne al tamaño de la muestra. Como se documentó en el Capítulo 3, los 54 participantes satisfacen los requerimientos de la fase exploratoria, pero no los de una calibración paramétrica bajo la psicometría TRI, que exige muestras superiores a 200 casos. Los estadísticos reportados en el Capítulo 4 son, por tanto, estimaciones orientadoras cuya estabilidad deberá confirmarse en la siguiente fase con un volumen de datos mayor.

La segunda condición atañe al formato del instrumento. Una prueba de selección múltiple evalúa el conocimiento declarativo y procedimental teórico, pero no la ejecución práctica del diseño. La distancia entre saber identificar un criterio de licencia y aplicarlo correctamente en un proyecto real, o entre reconocer la eficacia comunicativa de una imagen y producir una con esa eficacia, no es capturada por este tipo de instrumento. Esta limitación es inherente al formato de la prueba, no corregible mediante ajustes al banco de ítems, y debe tenerse en cuenta al interpretar los resultados como diagnóstico del perfil de egreso.

La tercera condición es el contexto de aplicación. La administración digital no controlada pudo haber introducido variables exógenas —consultas externas, interrupciones, falta de condiciones de evaluación estandarizadas— que influyeron en los puntajes y contribuyeron parcialmente al efecto techo observado. Una aplicación presencial con condiciones controladas permitiría deslindar con mayor precisión cuánto del efecto techo corresponde al nivel real de competencia de los estudiantes y cuánto a las condiciones de la prueba.

### 5.4. Ruta de desarrollo

Los hallazgos de esta investigación permiten trazar con precisión los pasos que el instrumento requiere para consolidarse como herramienta de uso institucional. Esa ruta tiene tres horizontes que responden a lógicas distintas y que conviene distinguir.

El trabajo inmediato es de depuración técnica. Los ítems 1 y 20 deben retirarse definitivamente del banco de preguntas por su discriminación negativa, y los ítems 6, 14, 19 y 30 requieren reformulación de distractores y enunciados para recuperar su capacidad discriminativa. Paralelamente, el efecto techo identificado en el piloto exige incorporar al menos diez reactivos de alta complejidad hermenéutica anclados en situaciones visuales inéditas o en problemas que el currículo no aborde explícitamente, de manera que el instrumento pueda discriminar con precisión en los niveles superiores de habilidad donde actualmente pierde resolución.

El horizonte de mediano plazo es la calibración definitiva. Con la versión depurada del instrumento y una muestra ampliada a más de 200 participantes, será posible realizar el análisis formal de ajuste al Modelo de Rasch mediante software especializado, establecer índices estandarizados y construir mapas de variables definitivos que permitan la equiparación de puntuaciones en aplicaciones futuras. Esta fase transformará el prototipo validado en un instrumento calibrado, con una escala en logits que ordene a los estudiantes de manera comparable entre cohortes y entre programas.

El horizonte de largo plazo es la implementación institucional. La aplicación de la prueba en dos momentos de la carrera —a mitad y al final del programa— permitiría medir el valor agregado del currículo en el desarrollo de la alfabetización visual, convirtiendo el instrumento en un insumo para decisiones pedagógicas fundamentadas. En su versión más desarrollada, el banco de ítems podría transformarse en una prueba adaptativa computarizada que ajuste su nivel de exigencia al desempeño de cada estudiante, optimizando la precisión de la medida con un número menor de reactivos.

Esta investigación partió de una carencia concreta: el programa de Diseño Digital y Multimedia de la Universidad Colegio Mayor de Cundinamarca no contaba con un instrumento

que permitiera saber, con criterio estandarizado, si sus estudiantes alcanzaban los niveles de alfabetización visual correspondientes al perfil de egreso. El prototipo desarrollado aquí no resuelve ese problema de forma definitiva, pues requiere una muestra mayor y una calibración Rasch completa para consolidarse como herramienta institucional, pero demuestra que el problema tiene solución técnica. Los datos del piloto revelan que los egresados del programa dominan las competencias visuales técnicas y éticas, y que las hermenéuticas presentan mayor variabilidad. Esa información, por sí sola, ya orienta decisiones curriculares. La pregunta que queda abierta para la siguiente fase es si esa variabilidad en la lectura crítica es un efecto del instrumento —ítems que requieren ajuste— o un efecto del currículo: una dimensión que se enseña de forma menos sistemática que las otras. Responder esa pregunta requiere el instrumento calibrado que esta investigación deja preparado.

## LISTA DE REFERENCIAS

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Association of College and Research Libraries. (2011). Visual literacy competency standards for higher education. American Library Association. https://www.ala.org/acrl/standards/visualliteracy (Acceso Abril 20, 2026)

Association of College and Research Libraries. (2022). Marco de la ACRL para la alfabetización informacional en la educación superior: alfabetización visual. https://www.ala.org/sites/default/files/acrl/content/standards/Framework_Companion_Visual_Literacy.pdf

Avgerinou, M. D., & Pettersson, R. (2011). Toward a cohesive theory of visual literacy. *Journal of Visual Literacy*, 30(2), 1-19.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Börner, K., Bueckle, A., & Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. Proceedings of the National Academy of Sciences, 116(6), 1857-1864. https://doi.org/10.1073/pnas.1807180116

Brumberger, E. (2025). (Re)defining visual literacy for the age of generative artificial intelligence. Journal of Visual Literacy, 44(4), 1–17. https://doi.org/10.1080/1051144X.2025.2566593

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education (*8th ed.). Routledge.

Cui, Y., Ge, L. W., Ding, Y., Yang, F., Harrison, L., & Kay, M. (2023). Adaptive assessment of visualization literacy. arXiv. https://doi.org/10.48550/arXiv.2308.14147

Debes, J. L. (1969). The loom of visual literacy. Audiovisual Instruction, 14(8), 25-27.

Denzin, N. K. (2009). The research act: A theoretical introduction to sociological methods. Aldine Transaction. (Obra original publicada en 1978). https://archive.org/details/researchacttheo00denz/page/n1/mode/2up

Escobar-Pérez, J., & Cuervo-Martínez, A. (2008). Validez de contenido y juicio de expertos: una aproximación a su utilización. *Avances en Medición,* 6(1), 27-36. https://www.researchgate.net/publication/302438451_Validez_de_contenido_y_juicio_de_expertos_Una_aproximacion_a_su_utilizacion

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Allyn & Bacon.

Hattwig, D., Bussert, K., Medaille, A., & Burgess, J. (2013). Visual literacy standards in higher education: New opportunities for libraries and student learning. portal: Libraries and the Academy, 13(1), 61-89. https://doi.org/10.1353/pla.2013.0008

Hernández-Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación* (6.ª ed.). McGraw-Hill Education.

Hill, R. L., & Conceição, S. C. O. (2025). Media literacy and faculty responsibility: Addressing AI-generated disinformation in higher education. New Directions for Adult and Continuing Education, 2025(188), 63–69. https://doi.org/10.1002/ace.70018

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size,

communality, and overdetermination. Educational and Psychological Measurement, 65(2), 202-226. https://doi.org/10.1177/0013164404267287

ICFES. (2018). Saber al detalle (Edición 13). Instituto Colombiano para la Evaluación de la Educación. https://www.icfes.gov.co/wp-content/uploads/2024/11/05102023-Saber-al-Detalle-ED-13.pdf

ICFES. (2019). Marco de referencia del módulo generación de artefactos - Saber Pro. Instituto Colombiano para la Evaluación de la Educación. https://www.icfes.gov.co/wp-content/uploads/2024/11/MR-Generacion-de-artefactos-Saber-Pro.pdf

Kane, M. T. (2013). Validating the interpretations and uses of test scores. J*ournal of Educational Measuremen*t, 50(1), 1-73. https://doi.org/10.1111/jedm.12000

Kędra, J. (2018). What does it mean to be visually literate? *Journal of Visual Literacy,* 37(2), 67-84. https://doi.org/10.1080/1051144X.2018.1492234

Lee, S., & Kwon, B. C. (2017). VLAT: Development of a visualization literacy assessment test. *IEEE Transactions on Visualization and Computer Graphics*, 23(1), 551-560. https://doi.org/10.1109/TVCG.2016.2598920

Li, J., & Potter, K. (2023). Visual literacy assessment: A review. *Information and Learning Sciences*, 124(7/8), 689-703. https://doi.org/10.1108/ILS-03-2023-0037

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4), 328.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. Measurement: *Interdisciplinary Research and Perspectives*, 1(1), 3-62.

Morales, H. (2022). Alfabetización visual en Colombia: desafíos para la educación superior. *Revista Colombiana de Educación*, 12(1), 45-60.

Müller, M. G. (2008). Visual competence: A new paradigm for studying visuals in the social sciences. *Visual Studies*, 23(2), 101-112. https://doi.org/10.1080/14725860802276248

Muñiz, J. (2010). Las teorías de los tests: Teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo*, 31(1), 57-66.

Natharius, D. (2004). The more we know, the more we see: The role of visuality in media literacy. *American Behavioral Scientist*, 48(2), 238-247. https://doi.org/10.1177/0002764204267253

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Pandey, A. V., & Ottley, A. (2023). Mini-VLAT: A short visual literacy test. *Computer Graphics Forum*, 42(3), 455-467. https://doi.org/10.1111/cgf.14809

Thomson, T. J., Pfurtscheller, D., Lobinger, K., Laba, N., & Christ, K. (2025). Visual literacies in transition: Multimodal co-production with generative AI. Journal of Visual Literacy, 44(4). https://doi.org/10.1080/1051144X.2025.2566592

Thompson, J., & Beene, S. (2021). Uniting the field of visual literacy: A call to action. *Journal of Visual Literacy*, 40(1), 1-11. https://doi.org/10.1080/1051144X.2021.1883152

Universidad Colegio Mayor de Cundinamarca. (2024). Proyecto Educativo del Programa de Diseño Digital y Multimedia. https://www.universidadmayor.edu.co/atencion-servicios-ciudadania/programas/pregrados/diseno-digital-multimedia/folleto-informacion-general/proyecto-educativo-del-programa-pep.

# ANEXOS

Anexo A. Estándares ACRL con indicadores de desempeño y resultados de aprendizaje.

Anexo B. Marco de especificaciones DCE: afirmaciones, evidencias y tareas.

Anexo C. Especificaciones técnicas del instrumento de medición.

Anexo D. Protocolo y registro de generación de estímulos visuales mediante IA.

Anexo E. Cuadernillo de validación para panel de jueces expertos.

Anexo F. Observaciones del panel de jueces evaluadores.