PROCEDIMIENTO ARTICULADO DE EVALUACIÓN EX-ANTE PARA ARTEFACTOS DESIGN SCIENCE RESEARCH

1er M.Sc. Paul Fernando Grimaldo Bravo
Posgrado SOE – UAGRM
https://orcid.org/0009-0000-6343-9684
Santa Cruz, Bolivia | pgrimaldo@stc.soeuagrm.edu.bo

2do PhD. Luis Roberto Pérez Rios
Posgrado SOE – UAGRM
https://orcid.org/0000-0002-8385-1016
Santa Cruz, Bolivia | luis.roberto@alenasoft.com

https://doi.org/10.23670 FT.2026.1.38

Recibido 08/05/2026 - Aceptado 29/05/2026

RESUMEN

La evaluación constituye una actividad nuclear del paradigma Design Science Research (DSR) y una de las que menos guías operacionales ha recibido en la literatura, de manera particular cuando el artefacto evaluado es una guía, un framework o un modelo de proceso que no admite experimentación controlada. El marco FEDS (Framework for Evaluation in Design Science) aporta la conceptualización evaluativa, pero no prescribe cómo seleccionar evaluadores con criterios objetivos, cómo estructurar el instrumento ni qué estadísticos aplicar según la estructura de los datos. Existe, en paralelo, un repertorio consolidado de métodos cuantitativos para validación por juicio de expertos que combina aportes de la tradición iberoamericana (coeficiente de competencia K, W de Kendall, puntos de corte de Torgerson) con desarrollos de validez de contenido de circulación internacional (CVC de Hernández-Nieto, V de Aiken), aunque rara vez aparece articulado con los marcos DSR. Ante este escenario, el presente artículo propone un procedimiento de siete fases para la evaluación ex-ante de artefactos DSR no experimentables, ubicado en el cuadrante artificial ex-ante de FEDS, cuya fase de análisis habilita la selección del instrumento estadístico en función del tipo de juicio perseguido y de la estructura ordinal de los datos. El trabajo se apoya en revisión documental, articulación conceptual entre las tradiciones y verificación interna frente a criterios de operatividad, replicabilidad, trazabilidad e integración teórica declarados ex-ante. Se concluye que el procedimiento se propone como una contribución metodológica operacional aplicable por investigadores que evalúan artefactos DSR no experimentables.

Palabras clave: Design Science Research; evaluación ex-ante; FEDS; validación por juicio de expertos; validez de contenido.

ABSTRACT

Evaluation is a core activity of the Design Science Research (DSR) paradigm and one of the least supported by operational guidance in the literature, particularly when the evaluated artifact is a guideline, a framework or a process model that does not allow controlled experimentation. The FEDS framework (Framework for Evaluation in Design Science) provides a consolidated evaluative conceptualization, but does not prescribe how to select evaluators under objective criteria, how to structure the consultation instrument, or which statistical tools to apply depending on the structure of the data. In parallel, there exists a consolidated repertoire of quantitative methods for expert-judgment validation that combines contributions from the Ibero-American tradition (competence coefficient K, Kendall's W, Torgerson cut-points) with internationally circulating content-validity developments (Hernández-Nieto's CVC, Aiken's V), though it rarely appears articulated with reference DSR frameworks. In response, this article proposes an articulated seven-phase procedure for the ex-ante evaluation of non-experimentable DSR artifacts, positioned within the artificial ex-ante quadrant of FEDS, whose analytical phase enables the selection of the statistical instrument according to the type of judgment sought and the ordinal structure of the data. The work rests on a documentary review, a conceptual articulation between the traditions, and an internal verification of the procedure against criteria of operability, replicability, traceability and theoretical integration declared ex-ante. It is concluded that the procedure is proposed as an operational methodological contribution applicable by researchers evaluating non-experimentable DSR artifacts.

Keywords: Design Science Research; ex-ante evaluation; FEDS; expert judgment validation; content validity.

INTRODUCCIÓN

Hoy, el paradigma de Design Science Research (DSR) se encuentra consolidado como marco legítimo para la producción de conocimiento en ingeniería de software, sistemas de información y disciplinas afines, particularmente cuando el resultado de la investigación adopta la forma de un artefacto prescriptivo: modelo, método, guía o instanciación orientada a resolver problemas organizacionales reales (Hevner et al., 2004; Peffers et al., 2007). En este paradigma, la evaluación no es una etapa opcional ni posterior al cierre del trabajo; ya March y Smith (1995) la incorporaron como una de las cuatro actividades constitutivas de la investigación en diseño, junto con la construcción, la teorización y la justificación, sentando con ello la base conceptual sobre la que descansan los desarrollos posteriores del paradigma. Es la actividad que provee retroalimentación al ciclo de diseño y, cuando se ejecuta con rigor, asegura la calidad científica del artefacto producido (Venable et al., 2016). La centralidad de la evaluación en DSR no es discutida en lo conceptual, pero en la práctica presenta dificultades operacionales que la literatura no siempreexplicita.

Un porcentaje sustantivo de los artefactos que se producen en DSR, en particular las guías metodológicas, los frameworks de referencia, los modelos de proceso y las recomendaciones de ingeniería, no admite experimentación controlada en el sentido estricto. Se considera no experimentable a aquel artefacto cuya naturaleza prescriptiva o metodológica imposibilita la asignación aleatoria de unidades organizacionales a condiciones de tratamiento y control, o cuya instanciación previa en condiciones equivalentes al entorno real de aplicación resulta inviable por restricciones de tiempo, acceso o escala. Ante este escenario, la evaluación ex-ante mediante juicio de expertos se ha consolidado como alternativa metodológicamente apropiada (Venable et al., 2016; Prat et al., 2015). La pregunta relevante no es si el juicio experto constituye una opción válida, dado que eso ya está establecido, sino cómo operacionalizarlo con rigor.

El marco FEDS, propuesto por Venable et al. (2016), constituye la referencia más citada en la literatura DSR para orientar la selección estratégica del paradigma evaluativo en función del propósito y el contexto de la evaluación. Se articula en dos dimensiones, propósito funcional (formativo o sumativo) y paradigma evaluativo (artificial o naturalista), y define cuatro cuadrantes entre los que el investigador debe ubicar cada episodio evaluativo. El cuadrante artificial ex-ante aplica de manera particular a la evaluación anticipada de artefactos antes de su instanciación en un entorno real, y es allí donde se ubican naturalmente los artefactos metodológicos no experimentables. No obstante, FEDS se detiene en la conceptualización estratégica y no prescribe procedimientos operacionales: no indica cómo seleccionar evaluadores con criterios objetivos, cómo estructurar el instrumento de consulta, cómo procesar estadísticamente las respuestas ordinales ni cómo traducir la concordancia observada en una decisión evaluativa trazable. Esta es una limitación reconocida incluso por autores que adoptan FEDS como referencia central (Sonnenberg y vom Brocke, 2012; Prat et al., 2015).

En paralelo a la evolución de los marcos DSR, para la validación por juicio de expertos se ha consolidado un repertorio de métodos cuantitativos con instrumentos específicos para distintos objetivos evaluativos. De la tradición iberoamericana y la psicometría internacional provienen instrumentos como el coeficiente de competencia K (Ramírez-Urizarri, 1999), el coeficiente de concordancia W de Kendall (Alvarado, 2008), la técnica de puntos de corte de Torgerson (Torgerson, 1958), el coeficiente de validez de contenido CVC (Hernández-Nieto, 2002) y el coeficiente V de Aiken (Aiken, 1985; Penfield y Giacobbi, 2004). Aunque estos instrumentos aparecen aplicados en revistas indexadas en dominios tan diversos como la educación, la salud, la ingeniería y los sistemas de información (Escobar-Pérez y Cuervo-Martínez, 2008; Cabero y Llorente, 2013; Galicia Alarcón et al., 2017; Marín-González et al., 2021;Herrera-Masó et al., 2022; Pedrosa et al., 2014), rara vez se los presenta como un repertorio único articulado con un paradigma DSR de circulación internacional. El resultado es una desarticulación que limita tanto la operatividad de FEDS como la circulación internacional de un aparato metodológico maduro.

La fundamentación teórica del presente trabajo se sostiene en tres cuerpos de literatura convergentes. El primero es el paradigma DSR en su versión canónica (Hevner et al., 2004; Peffers et al., 2007) y sus desarrollos contemporáneos sobre tipos de contribución (Gregor y Hevner, 2013; Baskerville et al., 2018), que permiten ubicar aportes metodológicos al propio proceso DSR como contribuciones de pleno derecho al campo. El segundo es la literatura específica sobre evaluación en DSR (Sonnenberg y vom Brocke, 2012; Prat et al., 2015; Venable et al., 2016), que aporta la conceptualización del cuadrante evaluativo y los criterios de rigor aplicables. El tercero es la literatura consolidada sobre métodos cuantitativos para validación por juicio de expertos, con dos vertientes conectadas: la tradición iberoamericana de consulta a expertos (Ramírez-Urizarri, 1999; Alvarado, 2008; Escobar-Pérez y Cuervo-Martínez, 2008; Cabero y Llorente, 2013) y la literatura internacional sobre validez de contenido en psicometría (Aiken, 1985; Hernández-Nieto, 2002; Penfield y Giacobbi, 2004; Pedrosa, Suárez-Álvarez y García-Cueto, 2014). La coherencia entre los tres cuerpos descansa en un supuesto epistemológico compartido, el pragmatismo de la ciencia del diseño, y se materializa en que todos los enfoques buscan, por vías distintas, producir juicios evaluativos rigurosos sobre artefactos que no admiten experimentación. En esta misma línea, conviene anticipar la ubicación tipológica del presente aporte dentro del cuadrante de contribuciones de Gregor y Hevner (2013). El trabajo se sitúa de manera predominante en la categoría de Improvement, en tanto aborda un problema reconocido en un dominio consolidado mediante una solución cuya forma articulada no había sido formalizada con

anterioridad. No se proponen instrumentos estadísticos nuevos, sino una orquestación procedimental que mejora la operatividad del marco FEDS al integrar de manera trazable un repertorio cuantitativo cuya pertinencia individual ya estaba demostrada en otras tradiciones.

La justificación del trabajo se articula en tres planos. En el plano académico, el trabajo aporta un procedimiento operacional que cubre vacíos reconocidos en la literatura de evaluación DSR, lo cual amplía el repertorio metodológico disponible para investigadores que trabajan con artefactos no experimentables. En el plano metodológico, el trabajo tiende un puente entre la literatura DSR de referencia internacional y un repertorio de métodos cuantitativos para validación por juicio de expertos que ha evolucionado, hasta ahora, en paralelo y sin contacto sustantivo con dicha literatura. En el plano práctico, el procedimiento propuesto ofrece a investigadores un instrumento replicable y trazable que reduce la improvisación metodológica y formaliza decisiones habitualmente tomadas con base en criterios subjetivos.

A partir de lo expuesto, el objetivo general del presente trabajo es proponer un procedimiento articulado de evaluación ex-ante aplicable a artefactos Design Science Research no experimentables, que integre el marco FEDS con el repertorio consolidado de métodos cuantitativos para validación por juicio de expertos. De este objetivo se derivan tres objetivos específicos: (1) identificar los vacíos operacionales del marco FEDS en el cuadrante artificial ex-ante; (2) sistematizar el repertorio de métodos cuantitativos para validación por juicio de expertos en términos compatibles con las categorías de FEDS, explicitando la relación entre el tipo de juicio perseguido y el instrumento estadístico apropiado; y (3) articular un procedimiento operacional estructurado en fases, con entradas, actividades y salidas trazables, que cubra los vacíos identificados.

DESARROLLO

El desarrollo de esta reflexión se fundamenta en un análisis crítico de las limitaciones operacionales del marco FEDS, específicamente en su aplicación a artefactos que no admiten experimentación controlada. Mediante la integración de la literatura psicométrica internacional y los métodos de consulta a expertos de la tradición iberoamericana, se articula una propuesta procedimental que busca formalizar la evaluación ex-ante. Este análisis permite transitar desde la conceptualización estratégica hacia una operacionalización estadística trazable, asegurando que la validación del artefacto responda a criterios de rigor y pertinencia metodológica.

Articulación del repertorio cuantitativo con los vacíos de FEDS

La superación de los vacíos operacionales de FEDS requiere la integración de instrumentos que respondan a la naturaleza del juicio evaluativo perseguido. Para el vacío de selección de evaluadores, se adopta el coeficiente de competencia K (Ramírez-Urizarri, 1999), cuya expresión es $K = 0{,}5(K_c + K_a)$, donde Kc es el coeficiente de conocimiento derivado de la autovaloración del experto en una escala de diez niveles y K_a el coeficiente de argumentación, ponderado a partir de la diversidad y solidez de sus fuentes de fundamentación teórica y práctica; únicamente los candidatos con $K \geq 0{,}7$ (nivel medio) o $K \geq 0{,}9$ (nivel alto) integran el panel definitivo.

Respecto al procesamiento estadístico, el repertorio se selecciona según el objetivo de la consulta. Cuando se busca medir el consenso en la jerarquización de atributos, se emplea el coeficiente **W** de Kendall con su prueba de x^2 asociada (Alvarado, 2008; Siegel y Castellan, 1988), cuya expresión con corrección por empates es:

$$W = \frac{12S}{K^2(N^3 - N) - K\sum_j T_j}$$

donde S es la suma de los cuadrados de las desviaciones de los rangos respecto a su media global, K el número de evaluadores, N el número de ítems ordenados, y $\sum_j T_j$ la corrección por empates calculada como $T_j = \sum_t (t^3 - t)$ para cada grupo de t rangos iguales del evaluador j; la significancia se contrasta mediante $x^2 = K(N - 1)W$ con $N - 1$ grados de libertad, rechazando la hipótesis nula de ausencia de concordancia cuando dicho valor supera el umbral crítico.

Por el contrario, si el objetivo es validar la adecuación absoluta de los aspectos del artefacto mediante escalas Likert, se utilizan el Coeficiente de Validez de Contenido (CVC) de Hernández- Nieto (2002), con umbral de aceptación de 0,80, y el coeficiente V de Aiken (Aiken, 1985; Penfield y Giacobbi, 2004). El CVC de cada ítem i se calcula como $CVC_i = E_i/V_{max} - P_e$, donde E_i es la media de las valoraciones de los n expertos, V_{max} el valor máximo de la escala y $P_e = (1/V_{max})^n$ la corrección por probabilidad de error aleatorio; el CVC global es la media aritmética de los CVC_i individuales. Se adopta esta formulación simplificada de Hernández-Nieto (2002) frente a la variante que aplica una corrección adicional por probabilidad de error inter-juez antes del promedio; la elección entre ambas debe declararse explícitamente en cada estudio que invoque el procedimiento, dado que en paneles pequeños la convergencia entre formulaciones no está garantizada.

El coeficiente V, a su vez, se obtiene como $V = S/[n(c - 1)]$, donde S es la sumatoria de las diferencias entre cada valoración y el mínimo de la escala, n el número de expertos y c el número de categorías; su intervalo de confianza al 95 % se estima a partir de la distribución binomial exacta (Penfield y Giacobbi, 2004), aceptando el ítem en consulta única cuando el límite inferior del intervalo de confianza supera 0,50, criterio que opera como umbral mínimo de aceptación individual. Estos dos coeficientes miden planos complementarios– adecuación promedio y significancia estadística de esa adecuación– y deben leerse de forma conjunta.

Finalmente, para el juicio evaluativo integrado, se aplica

la técnica de puntos de corte de Torgerson (1958), cuya ejecución sobre la distribución de respuestas de cada ítem sigue cuatro pasos: (1) calcular las frecuencias relativas por categoría; (2) obtener las frecuencias relativas acumuladas; (3) convertir cada acumulado al valor *x* correspondiente en la distribución normal estándar mediante $x = \Phi^{-1}(T)$; y (4) derivar los puntos de corte entre categorías adyacentes como la media de los valores *x* contiguos, asignando a cada ítem la etiqueta cualitativa —Muy Adecuado, Bastante Adecuado, Adecuado, Poco Adecuado o Nada Adecuado— del intervalo en que cae su frecuencia acumulada modal, lo que traduce la distribución estadística en una decisión evaluativa directamente accionable. La técnica descansa en el supuesto de que el atributo evaluativo subyacente se distribuye normalmente sobre los expertos; el supuesto es razonable en paneles moderados, pero pierde tracción cuando *n* es pequeño o cuando las respuestas se concentran en una única categoría, situaciones en las que los puntos de corte deben interpretarse como aproximaciones orientativas más que como umbrales estrictos. En la Tabla 1 se evidencia la articulación de los vacíos de FEDS con los instrumentos y estructuras de datos.

Tabla 1

Matriz de articulación: Vacíos de FEDS, instrumentos y estructura de datos en el cuadrante artificial ex-ante

Vacío operacional de FEDS	Pregunta evaluativa sobre el artefacto	Estructura de datos	Instrumento principal	Fase del procedimiento
Selección de experto	¿Es el candidato competente en el dominio?	Autovaloración y fuentes	Coeficiente K (Kc + Ka)	Fase 3
Diseño del instrumento	¿Qué aspectos del artefacto se deben evaluar?	Matriz de dimensiones del artefacto	Escala Likert o Rangos puros	Fase 2 y 4
Procesamiento estadístico	¿Los expertos valoran positivamente cada aspecto?	Likert ordinal (valoración absoluta)	CVC + V de Aiken (IC 95%)	Fase 6
	¿Los expertos coinciden en la jerarquía de los ítems?	Rangos ordinales o transformación a rangos	W de Kendall + X^2	Fase 6
Juicio integrado	¿En qué categoría evaluativa cae cada ítem?	Frecuencias relativas acumuladas	Puntos de corte de Torgerson	Fase 7

Nota. *Elaboración propia a partir de Aiken (1985), Alvarado (2008), Hernández-Nieto (2002), Pedrosa et al. (2014), Ramírez-Urizarri (1999), Siegel y Castellan (1988) y Venable et al. (2016). La elección de W de Kendall para concordancia ordinal sobre alternativas como* α *de Krippendorff o AC1 de Gwet responde a su arraigo en la tradición iberoamericana de evaluación por expertos y a la disponibilidad de su prueba* X^2 *asociada; las alternativas se reseñan como línea de exploración futura en la sección de limitaciones.*

La identificación de estos vacíos no constituye una crítica al marco FEDS, concebido originalmente como una guía de orientación estratégica, sino que subraya la necesidad de un complemento operativo que evite la proliferación de procedimientos ad-hoc y garantice la trazabilidad entre estudios.

La sistematización propuesta integra instrumentos de la tradición iberoamericana con desarrollos internacionales de validez de contenido, asegurando que cada dimensión crítica, desde la selección objetiva de expertos mediante el coeficiente K hasta la categorización de resultados con la técnica de Torgerson, cuente con un respaldo metodológico robusto. Esta integración permite cubrir integralmente las demandas operacionales del cuadrante artificial ex-ante, transformando decisiones habitualmente subjetivas en procesos formalizados y replicables.

Como se detalla en la Tabla 1, la articulación de este repertorio no es mecánica, sino que depende de la correspondencia directa entre la pregunta evaluativa y la estructura de los datos. Mientras que el coeficiente CVC y la V de Aiken se especializan en cuantificar la valoración absoluta de los atributos del artefacto mediante escalas Likert, el coeficiente W de Kendall se reserva para medir el consenso en la jerarquización de los mismos, ajustando el análisis a la naturaleza ordinal de los rangos. Esta distinción conceptual asegura que la elección del instrumento estadístico en la Fase 6 del procedimiento sea una decisión trazable y declarada ex-ante, permitiendo que el investigador justifique el rigor de su juicio evaluativo en función de los objetivos específicos de la validación.

Procedimiento articulado propuesto

El procedimiento se estructura en siete fases secuenciales agrupadas en tres bloques funcionales: conceptualización evaluativa (FEDS, fases 1 y 2), operacionalización estadística (repertorio cuantitativo, fases 3 a 5) y análisis e interpretación integrada (fases 6 y 7).

El número de fases responde a una decisión de diseño expositivo orientada a maximizar la trazabilidad sin redundancia: estructuras más compactas fundirían decisiones que operan sobre fundamentos distintos —v.g., selección de evaluadores y diseño del instrumento, o análisis estadístico y decisión evaluativa—, mientras que estructuras más granulares introducirían subdivisiones operativamente equivalentes.

La Tabla 2 sintetiza las entradas, actividades y salidas de la estructura completa.

Tabla 2

Estructura del procedimiento articulado de evaluación ex-ante

Fase	Nombre	Entradas	Actividades principales	Salidas
1	Caracterización del artefacto y ubicación en FEDS	Artefacto DSR a evaluar	Clasificar por instanciabilidad, madurez y experimentabilidad; ubicar en cuadrante FEDS	Cuadrante evaluativo justificado
2	Formulación de dimensiones evaluativas	Artefacto + cuadrante	Derivar dimensiones; articular sub-criterios	Matriz de dimensiones y sub-criterios
3	Selección de expertos mediante coeficiente K	Pool de candidatos	Calcular Kc, Ka y K; filtrar por nivel medio o alto	Panel con coeficientes documentados
4	Diseño del instrumento de consulta	Matriz de dimensiones + tipo de juicio perseguido	Traducir dimensiones en ítems con escala Likert de cinco categorías o con rangos ordinales puros, según Tabla 1; incluir ítems abiertos opcionales	Cuestionario trazable a la matriz
5	Administración y recolección	Panel + cuestionario	Administración asincrónica, anónima, individual; ronda única o adicionales	Matriz de respuestas ordinales
6	Análisis estadístico de concordancia y cuantificación	Matriz de respuestas	Seleccionar instrumento del repertorio (W de Kendall con x^2 asociado; CVC; V de Aiken con IC 95 %) según Tabla 1; puntos de corte de Torgerson	Evidencia de concordancia + categorización por aspecto
7	Interpretación integrada y decisión evaluativa	Resultados estadísticos (+ comentarios abiertos opcionales)	Integrar evidencia cuantitativa; triangular con comentarios abiertos cuando estén disponibles; decidir validación, iteración o rediseño	Juicio evaluativo integrado

La ejecución de este flujo procedimental obedece a una lógica de trazabilidad estricta. En el primer bloque, la Fase 1 sitúa estratégicamente la evaluación justificando el cuadrante FEDS a partir de atributos estructurales como la instanciabilidad y madurez del artefacto según lo discutido anteriormente. Esta decisión condiciona la Fase 2, donde el artefacto se deconstruye en una matriz de dimensiones evaluables. Candidatos habituales son pertinencia teórica, suficiencia estructural, aplicabilidad operativa y coherencia interna, aunque la selección concreta depende del tipo de artefacto. Cada dimensión se descompone en sub-criterios evaluables por separado. La matriz orienta el diseño del instrumento en la Fase 4 y asegura cobertura balanceada de los aspectos del artefacto.

En el bloque de operacionalización, la Fase 3 supera la selección subjetiva tradicional al documentar matemáticamente la inclusión exclusiva de perfiles. El procedimiento convoca a candidatos que cumplan criterios previos de elegibilidad (formación académica, experiencia, vinculación al dominio), aplica un cuestionario de autovaloración y calcula K_c, K_a y K para cada uno; únicamente los candidatos con nivel medio o alto integran el panel. Se recomienda que el panel definitivo alcance al menos siete expertos elegibles, con un rango preferente entre ocho y quince integrantes para garantizar estabilidad estadística de los coeficientes ulteriores (Escobar-Pérez y Cuervo-Martínez, 2008; Cabero y Llorente, 2013). La Fase 4 consiste en la operacionalización de la matriz de dimensiones en un cuestionario estructurado compuesto por ítems cerrados de escala ordinal. La selección de la métrica no es arbitraria, sino que responde a la naturaleza del juicio requerido: se emplea una escala Likert de cinco niveles (desde Muy Adecuado hasta Nada Adecuado) para valoraciones de carácter absoluto —respaldada por los criterios de Aiken (1985), Alvarado (2008) y Cabero y Llorente (2013)— o, en su defecto, rangos puros *(1…k)* para tareas de jerarquización. Si bien es posible integrar de forma opcional ítems cualitativos abiertos, su alcance es limitado y funcional: actúan como un instrumento diagnóstico para justificar puntuaciones bajas y orientar el rediseño del artefacto. Es fundamental precisar que estos ítems no poseen una función validatoria per se, sino que sirven como insumo para la triangulación en la Fase 7. Durante la Fase 5, la administración asincrónica, anónima e individual del cuestionario resulta crítica para prevenir sesgos cognitivos inter-evaluadores, como el efecto líder —la tendencia documentada a que los evaluadores ajusten sus juicios hacia la posición del experto de mayor jerarquía o visibilidad en el panel— (Cabero-Almenara e Infante-Moro, 2014). La ronda única es la opción habitual; si no alcanza concordancia aceptable, se admiten rondas adicionales bajo lógica Delphi (Okoli y Pawlowski, 2004). El criterio de parada para estas iteraciones adicionales se determinará según la naturaleza del instrumento: (a) para artefactos que requieren evaluación ordinal o jerárquica, el ciclo concluye al alcanzar una concordancia de Kendall $W \geq 0,70$ con X^2 significativo $(p < 0,05)$ (Alvarado, 2008;Herrera-Masó et al., 2022); (b) para artefactos que requieren evaluación absoluta, se detendrá cuando los ítems superen un $CV\ V \geq 0,80$ (Hernández-Nieto, 2002) o, en su defecto, se obtenga una V de $Aiken \geq 0,78$ con todos los límites inferiores de los intervalos de confianza

sobre 0,50, criterio más exigente que el de aceptación individual descrito anteriormente, en tanto opera como umbral de convergencia tras iteración Delphi; o (c) el proceso finalizará mediante la convergencia declarada por el investigador bajo un criterio metodológico debidamente justificado (Okoli y Pawlowski, 2004).

Finalmente, en el bloque de cierre, la Fase 6 ejecuta el cálculo estadístico seleccionando el coeficiente pertinente (CVC, V de Aiken o W de Kendall) y aplicando los puntos de corte de Torgerson. La Fase 7 integra estos resultados cuantitativos —evidencia de concordancia estadística, valoración absoluta por ítem y categorización de Torgerson— en un juicio evaluativo con tres decisiones posibles: validación, iteración con ajustes o rediseño, este último en grado parcial o integral según la magnitud de los hallazgos.

Tabla 3

Matriz orientativa de decisión evaluativa en la Fase 7

Configuración de resultados	Decisión sugerida	Lectura interpretativa
CVC ≥ 0,80 (todos los ítems) y W ≥ 0,70 con X^2 significativo ($p < 0,05$); o V de Aiken ≥ 0,78 con LI > 0,50; categorización modal en "Bastante Adecuado" o "Muy Adecuado"	Validación	Concordancia y adecuación convergen positivamente
0,60 ≤ CVC < 0,80 en algunos ítems, o W entre 0,50 y 0,70, con categorización modal en "Adecuado"	Iteración con ajustes focalizados	Adecuación parcial o concordancia moderada que admite refinamiento sin rediseño estructural
CVC < 0,60 en una proporción sustantiva de ítems, W < 0,50, o categorización modal en "Poco Adecuado" o "Nada Adecuado"	Rediseño parcial o integral	Insuficiencia de adecuación o disenso estructural que excede el ajuste menor

Nota. Los umbrales se derivan de Hernández-Nieto (2002) para CVC, Alvarado (2008) y Herrera-Masó et al. (2022) para W de Kendall, y Penfield y Giacobbi (2004) para V de Aiken. La matriz opera como referencia orientativa, no como regla determinista.

La matriz de la Tabla 3 ofrece umbrales orientativos derivados de la literatura especializada para cada instrumento del repertorio, con el propósito de reducir la indefinición en la decisión final. No obstante, estos umbrales no operan como un algoritmo determinista: la naturaleza del artefacto, la dispersión de las valoraciones entre dimensiones y la evidencia cualitativa complementaria pueden justificar decisiones que se aparten de la lectura mecánica de la tabla. Se prescriben umbrales por economía operacional, pero la responsabilidad del juicio evaluativo final recae en el investigador, que debe declarar explícitamente la regla de decisión adoptada y sus eventuales desviaciones respecto de la matriz orientativa. Conviene señalar, en clave conceptual y no ya operativa, que la concordancia estadística mide coincidencia entre expertos y no adecuación intrínseca del artefacto, distinción que sustenta la lectura conjunta de los tres instrumentos del repertorio. Cuando el instrumento de consulta incluyó ítems abiertos, sus respuestas pueden triangularse de manera complementaria con los resultados cuantitativos, con el propósito acotado de identificar disonancias (por ejemplo, categoría alta con comentarios críticos) que orienten el rediseño del artefacto. Esta triangulación es opcional y no altera la naturaleza cuantitativa de la validación. Esta fase suele omitirse en reportes que cierran la validación con el solo resultado cuantitativo; su inclusión explícita refuerza la trazabilidad del juicio final.

Demostración del procedimiento

La verificación del procedimiento frente a los cuatro criterios declarados en la metodología arroja cumplimiento en los cuatro ejes: cada fase produce salidas específicas que alimentan la siguiente sin rupturas (operatividad); las siete fases están descritas con un nivel de detalle que permite su aplicación a artefactos distintos manteniendo el mismo esquema (replicabilidad); las salidas finales son trazables hasta las decisiones iniciales sin saltos no justificados (trazabilidad); y el procedimiento respeta los fundamentos conceptuales de FEDS y los supuestos de cada instrumento del repertorio considerando rangos ordinales para W de Kendall, Likert con valoración absoluta para CVC y V de Aiken, y frecuencias acumuladas para Torgerson, sin distorsionar ninguno para forzar la articulación (integración teórica).

Para evidenciar la ejecutabilidad del flujo en condiciones reales, el procedimiento fue aplicado por los autores a un artefacto metodológico del ámbito de la ingeniería de software en un trabajo previo, específicamente una guía metodológica con relevancia real en el área de arquitectura de software. La aplicación involucró un panel de 8 expertos con nivel de competencia medio o alto confirmado mediante coeficiente K, un cuestionario de 17 ítems cerrados y 2 preguntas abiertas opcionales derivado de una matriz de 8 dimensiones evaluativas (v.g., Aplicabilidad, Gestión de Riesgos) con escala Likert de cinco categorías (Nada Adecuado, Poco Adecuado, Adecuado, Bastante Adecuado, Muy Adecuado), y una ronda única de administración asincrónica y anónima. Dado que el objetivo evaluativo era validar positivamente cada aspecto del artefacto de forma absoluta, la Fase 6 adoptó CVC y V de Aiken como instrumentos principales, con Torgerson para la categorización final. Los 17 ítems superaron el umbral CVC ≥ 0,80 y los límites inferiores de los intervalos de confianza al 95 % de V de Aiken —calculados a partir de la distribución binomial exacta— se ubicaron por encima del criterio de aceptación individual establecido. El recorrido completo de las siete fases se ejecutó sin

rupturas operacionales, donde finalmente el artefacto en su totalidad alcanzó la categoría "Muy Adecuado" en la caracterización de Torgerson, lo cual respalda la aplicabilidad del procedimiento en condiciones reales. Los resultados numéricos detallados quedan fuera del alcance del presente artículo, cuyo objeto es el procedimiento y no el artefacto evaluado.

Alcance e interpretación del procedimiento propuesto

Los resultados obtenidos permiten interpretar el procedimiento propuesto como una contribución operacional al marco FEDS, no como un reemplazo de este. El marco conceptual de FEDS sigue siendo la referencia estratégica que guía la ubicación inicial de la evaluación y la decisión sobre el paradigma evaluativo a adoptar; el procedimiento aporta el detalle operativo. Ambos coexisten sin conflicto: FEDS para la estrategia, el procedimiento articulado para la ejecución.

Lo que efectivamente se gana con esta articulación no es un inventario adicional de instrumentos estadísticos, sino una regla de selección trazable. La experiencia en validación metodológica evidencia que el mayor costo evaluativo no recae en el cálculo matemático, sino en la indefinición previa de qué evaluar, por qué y con qué instrumento. En ausencia de un procedimiento formal, la inercia empuja al investigador a adoptar herramientas por mera tradición disciplinar, sin que la elección responda a la naturaleza del juicio perseguido. Al explicitar la correspondencia entre la pregunta evaluativa, la estructura de datos y el instrumento (Tabla 1), el procedimiento desplaza el esfuerzo desde la operatividad del cálculo hacia el rigor de la definición previa. Este enfoque atiende un nicho específico: la validación de artefactos metodológicos no experimentables, evaluados bajo restricciones críticas de tiempo y acceso que vuelven impracticable la instanciación previa o las múltiples rondas de convergencia. La articulación cuantitativa que aquí se propone no agota el espacio de evaluación ex-ante, sino que ofrece una formalización rigurosa para los casos donde la cuantificación del consenso es metodológicamente exigida.

Finalmente, de la Tabla 1 emerge una observación secundaria: los instrumentos presentados cubren los vacíos de FEDS, pero no agotan las posibilidades evaluativas. Rondas adicionales tipo Delphi, análisis de consistencia en la autovaloración o la adopción de coeficientes alternativos para paneles masivos podrían articularse en el futuro. En esa dirección, el procedimiento propuesto opera como una primera articulación mínima viable, abierta a enriquecimientos posteriores.

Limitaciones de la propuesta

El trabajo presenta cuatro limitaciones que deben considerarse al interpretar sus resultados. La primera es que la evaluación del procedimiento propuesto es de carácter interno y descansa principalmente en la verificación contra los criterios declarados ex-ante; la validación externa mediante su uso documentado por investigadores distintos a los autores, sobre una diversidad de artefactos y dominios, constituye la principal frontera del presente aporte. La segunda limitación es que la articulación propuesta se restringe al cuadrante artificial ex-ante de FEDS; los otros tres cuadrantes (naturalistic ex-ante, artificial ex-post y naturalistic ex-post) requieren procedimientos distintos que este trabajo no aborda. La tercera limitación, más sutil, es que el procedimiento hereda los supuestos estadísticos del repertorio cuantitativo incorporado: los criterios de interpretación, las reglas de corte y la lógica de selección plasmada en la Tabla 1 descansan en convenciones metodológicas consolidadas pero no exclusivas. Futuras versiones podrían explorar alternativas como el coeficiente alfa de Krippendorff, el kappa de Fleiss o modelos de teoría de respuesta al ítem.

Una cuarta limitación, de naturaleza epistemológica, atañe a la capacidad discriminativa del procedimiento. La aplicación previa que sostiene la verificación interna se ejecutó sobre un artefacto que superó los umbrales de aceptación, lo cual confirma la operatividad del flujo, pero no aporta evidencia sobre el comportamiento del procedimiento ante artefactos con defectos conocidos o estructuralmente débiles. Cabe la posibilidad de que los umbrales adoptados resulten poco sensibles ante casos límite o que la articulación de instrumentos genere falsos positivos en escenarios donde los expertos coinciden en valoraciones moderadas. La verificación de discriminabilidad requiere ejercicios deliberados de aplicación adversarial que excedan el alcance del presente artículo. Expresado en términos epistemológicos, la demostración prueba la operatividad del flujo —que ningún paso bloquea ni produce salidas inválidas— pero no su validez discriminativa, esto es, la capacidad del procedimiento para distinguir artefactos adecuados de artefactos deficientes.

Esta distinción entre operatividad demostrada y validez discriminativa pendiente debe asumirse al interpretar la verificación interna ofrecida y motiva la línea de trabajo futuro de aplicación adversarial. Existe además una tensión no resuelta entre la pretensión de generalización del procedimiento y las particularidades de cada artefacto DSR evaluado. El procedimiento prescribe las fases pero no puede prescribir las dimensiones evaluativas específicas, que dependen del tipo de artefacto y del dominio, lo cual deja al investigador con una decisión no trivial en la Fase 2. Esta tensión no invalida el procedimiento, pero delimita su alcance: es un marco operacional estructurado, no un algoritmo determinista.

CONCLUSIONES

El presente trabajo se propuso articular el marco FEDS con el repertorio consolidado de métodos cuantitativos para validación por juicio de expertos mediante un procedimiento operacional aplicable a la evaluación ex-ante de artefactos Design Science Research no experimentables. Los tres objetivos específicos enunciados en la introducción encuentran correspondencia directa con los resultados obtenidos: **Vacíos de FEDS identificados** (objetivo específico

1). Se caracterizaron cuatro vacíos operacionales del marco en el cuadrante artificial ex-ante (selección de evaluadores, diseño del instrumento, procesamiento estadístico y trazabilidad del juicio evaluativo integrado), cuya persistencia alimenta procedimientos ad-hoc no comparables entre estudios.

Repertorio cuantitativo sistematizado (objetivo específico 2). Se articuló el repertorio en términos compatibles con FEDS (coeficiente de competencia K, W de Kendall con X^2 asociado, puntos de corte de Torgerson, CVC de Hernández-Nieto y V de Aiken con intervalo de confianza) y se formalizó la correspondencia entre tipo de juicio evaluativo, estructura de datos e instrumento estadístico apropiado.

Procedimiento articulado construido (objetivo específico 3). Se estructuró un procedimiento de siete fases con entradas, actividades y salidas trazables que cubre los vacíos identificados, respeta los fundamentos conceptuales de FEDS y los supuestos estadísticos de cada instrumento, y quedó verificado frente a los cuatro criterios declarados ex-ante mediante una aplicación ilustrativa previa.

El objetivo general —proponer un procedimiento articulado aplicable a artefactos DSR no experimentables— se considera cumplido en su alcance propositivo, configurando una contribución de tipo Improvement, y quedando la validación externa por investigadores distintos a los autores como principal frontera del aporte. Como líneas de trabajo futuro se identifican tres direcciones. La primera es la validación externa del procedimiento mediante su uso documentado por investigadores distintos a los autores, sobre artefactos DSR de dominios diversos, con el fin de recoger evidencia empírica sobre su utilidad percibida, tiempo de aplicación, trazabilidad y necesidad de ajustes. La segunda es la extensión del procedimiento al cuadrante naturalistic ex-ante mediante la incorporación de focus groups estructurados y observación participante, de forma tal que el investigador pueda transitar entre cuadrantes FEDS con un único procedimiento articulado.

La tercera línea, de carácter más instrumental, consiste en el desarrollo de herramientas computacionales de soporte, tales como plantillas automatizadas para el cálculo de coeficientes, calculadoras para los puntos de corte de Torgerson e integraciones con plataformas de encuestas en línea, que reduzcan la barrera técnica de adopción del procedimiento. La cuarta línea, complementaria de la primera, consiste en un estudio de validación adversarial mediante la aplicación deliberada del procedimiento a artefactos con defectos conocidos o construidos ex profeso para someter a prueba su capacidad discriminativa. Este tipo de ejercicio permitiría verificar si los umbrales declarados detectan efectivamente los puntos débiles esperados o si requieren recalibración.

BIBLIOGRAFÍA

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. Educational and Psychological Measurement, 45(1), 131–142. https://doi.org/10.1177/0013164485451012

Alvarado, F. (2008). Análisis de concordancia de atributos. Tecnología en Marcha, 21(3), 29–35.

Baskerville, R., Baiyere, A., Gregor, S., Hevner, A., y Rossi, M. (2018). Design Science Research contributions: Finding a balance between artifact and theory. Journal of the Association for Information Systems, 19(5), 358–376. https://doi.org/10.17705/1jais.00495

Cabero, J., y Llorente, M. C. (2013). La aplicación del juicio de experto como técnica de evaluación de las tecnologías de la información y comunicación (TIC). Revista de Tecnología de Información y Comunicación en Educación, 7(2), 11–22.

Escobar-Pérez, J., y Cuervo-Martínez, Á. (2008). Validez de contenido y juicio de expertos: una aproximación a su utilización. Avances en Medición, 6(1), 27–36.

Galicia Alarcón, L. A., Balderrama Trápaga, J. A., y Edel Navarro, R. (2017). Validez de contenido por juicio de expertos: propuesta de una herramienta virtual. Apertura, 9(2), 42–53.

Gregor, S., y Hevner, A. R. (2013). Positioning and presenting Design Science Research for maximum impact. MIS Quarterly, 37(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

Hernández-Nieto, R. (2002). Contributions to statistical analysis. Universidad de Los Andes.

Herrera-Masó, J., Calero-Ricardo, J. L., González-Rangel, M. A., Collazo Ramos, M. I., y Travieso-González, Y. (2022). El método de consulta a expertos en tres niveles de validación. Revista Habanera de Ciencias Médicas, 21(1).

Hevner, A. R., March, S. T., Park, J., y Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105.

March, S. T., y Smith, G. F. (1995). Design and natural science research Decision Support Systems, 15(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

Marín-González, F., Paredes-Chacín, A. J., e Inciarte-González, A. (2021). Validación del diseño de una red de cooperación científico-tecnológica utilizando el coeficiente K para la selección de expertos. Información Tecnológica, 32(2), 79–88.

Okoli, C., y Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design considerations and applications. Information & Management, 42(1), 15–29. https://doi.org/10.1016/j.im.2003.11.0

Pedrosa, I., Suárez-Álvarez, J., y García-Cueto, E. (2014). Evidencias sobre la validez de contenido: avances teóricos y métodos para su estimación. Acción Psicológica, 10(2), 3–18. https://doi.org/10.5944/ap.10.2.11820

Peffers, K., Tuunanen, T., Rothenberger, M. A., y Chatterjee, S. (2007). A Design Science Research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Penfield, R. D., y Giacobbi, P. R. (2004). Applying a score confidence interval to Aiken's item content-relevance index. Measurement in Physical Education and Exercise Science, 8(4), 213–225. https://doi.org/10.1207/s15327841mpee0804_3

Prat, N., Comyn-Wattiau, I., y Akoka, J. (2015). A taxonomy of evaluation methods for information systems artifacts. Journal of Management Information Systems, 32(3), 229–267. https://doi.org/10.1080/07421222.2015.1099390

Ramírez-Urizarri, L. A. (1999). Algunas consideraciones acerca del método de evaluación utilizando el criterio de expertos. Instituto Superior de Ciencias Agropecuarias de La Habana.

Siegel, S., y Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences (2.ª ed.). McGraw-Hill.

Torgerson, W. S. (1958). Theory and methods of scaling. John Wiley and Sons.