Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Iztacala

"Evaluación Psicológica a través de un instrumento psicométrico digital"

T E S I S

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN PSICOLOGÍA

P R E S E N T A (N)

Ávalos Talavera Marco Antonio

Director: Lic. **José Manuel Sánchez Sordo**

Dictaminadores: Mtra. **Mirna Elizabeth Quezada**

Mtro. **J. Jesús Becerra Ramírez**

Los Reyes Iztacala, Edo de México, 2019

INDICE

INTRODUCCIÓN

La psicología se define generalmente como la ciencia de la conducta humana (Dunnette y Kirchner, 1989, Spector, 2002, Barbero, Vila y Holgado, 2010, Aamodt, 2010), misma que hace alusión a aquella actividad que realiza una o varias personas. Y es a través de ésta que podemos obtener implicaciones psicológicas inmediatas concernientes a su posición, método y contenido. Como ciencia, ésta intenta descubrir o desarrollar conceptos explicativos y métodos sistemáticos para lograr el control de su objeto de estudio; el hombre. En lo que va de la explicación, resulta evidente la exigencia de la identificación, descripción y observación de variables. Estas últimas deben estar sujetas a confirmación o invalidación por otros científicos que deseen repetir cualquier observación determinada. De tal manera que las opiniones, fantasías, argumentaciones, nociones favoritas o suposiciones no constituyen una fuente adecuada de conclusiones científicas; son los hechos revelados por procedimientos determinados de observación y experimentación los que fungen como fuente primordial de conclusiones científicas; en otras palabras, por medio de un método científico sistemático (Dunnette y Kirchner, 1989).

Dado lo anterior la psicología no trabaja en el reino de la fantasía acerca de la conducta humana; trabaja con hechos. Mismos que son sometidos a valoraciones; mediciones; análisis y modificaciones (en caso de ser posible) con el fin de propiciar un cambio significativo para la persona, grupo, comunidad, entidad o sociedad en la que se suscitó el evento. Ahora bien, para captar o recopilar información de estos hechos la psicología emplea la psicometría.

Barbero, Vila y Holgado (2010) definen a la psicometría como una disciplina metodológica dentro del área de la psicología cuya tarea fundamental es la medición o cuantificación de las variables psicológicas con todas las implicaciones que ello conlleva, tanto teóricas (posibilidades y criterios de medición) como prácticas (cómo y con qué se mide). Afirmando que la psicometría se encarga en primer lugar de la justificación y legitimación de la medición psicológica a través de: 1) el desarrollo de modelos formales que permitan representar los fenómenos que se quieren estudiar y posibiliten la

transformación de los hechos en datos; 2) la validación de los modelos desarrollados para determinar en qué medida representan la realidad que pretenden evaluar y, 3) estableciendo las condiciones que permitan llevar a cabo el proceso de medición pertinente. Y, en segundo lugar, se encarga de las implicaciones prácticas y aplicadas de la medición, es decir, proporciona los métodos necesarios sobre cómo se deben llevar a cabo las cuantificaciones de cada caso con base a la construcción de instrumentos necesarios y adecuados para tal labor. Por tal motivo, la psicometría se ve inmersa en diversos ámbitos psicológicos, tales como: educativo, organizacional, salud, político, social y cultural abarcando todos los marcos de la psicología: Personalidad, procesos cognitivos, actitudes, aptitudes, etc.

En lo que al área organizacional respecta, podemos destacar que desde hace muchos años atrás se lleva a cabo la aplicación de test para la selección y reclutamiento del personal, seguimiento de labor, capacitación, etc. Así mismo, se plantea, que las pruebas que principalmente se aplican en este ámbito son proyectiva, las cuáles son conocidas por solicitar al participante alguna actividad gráfica y/o verbal a partir de la percepción de un estímulo (Aamodt, 2010). Dichos instrumentos se realizan bajo el supuesto teórico de que cuanto menos estructurada resulte la tarea, menor control habrá en las respuestas por parte del participante, denotando su verdadera estructura psíquica (Muñiz, 2010). Sin embargo, con la evolución de la psicometría se obtuvo que dichas pruebas solían ser sumamente subjetivas en lo que, a la validez, confiabilidad, relación de pertinencia situacional de interpretación, generalización, errores de aplicación, carácter abiertos de los protocolos y subjetividad por parte del aplicador respecta (Brody, 1972, Erdelyi y Goldberg, 1979; Kinslinger, 1966, citados en Liporace y Solano, 2015). Volviendo de esta manera a las pruebas proyectivas una herramienta de dudosa capacidad evaluativa.

Con respecto a esto, Blum (1985) expone que una prueba es efectiva en el proceso que se pretende realizar cuando ésta cuenta con su respectiva ‘‘estandarización’’ poblacional, lo cual, a su vez permite obtener mediciones puntuales de la conducta de una muestra promedio en particular. Es decir, dicha muestra de conducta debe ser suficientemente grande y representativa del repertorio de comportamientos que se mida en una población específica para poder generalizar y predecir los resultados de las pruebas

consecuentes. En segundo lugar, al hablar de una prueba estandarizada también hace alusión a que el instrumento como tal cuenta con el nivel pertinente de validez, confiabilidad y evaluación promedio-normal de una muestra poblacional representativa; concibiendo de esta manera a una "norma" como un estándar de referencia; es decir, que nos permite comprender el significado de la puntuación en la prueba.

Existen dos sistemas que nos posibilitan valorar dicha puntuación: los baremos y las calificaciones estándar; ambos miden indirectamente la información dada sobre el desempeño de los individuos en la prueba en relación a la población conocida y nos habilita para poder ubicar a una persona en relación al grupo o a un todo (Prieto y Delgado, 2010, Nunnally, 1987; Magnusson, 1977). Sin embargo, las pruebas estandarizadas también presentan problemas de significación, "... una norma sólo será significativa cuando se conozcan las características de la población en que se basan" (Blum, 1985, pp. 131).

Muchos de los problemas con los que se enfrentan las pruebas psicométricas estandarizadas es su sensibilidad en tanto a su significación de puntuaciones a causa de la rigidez con la cual se construyeron. Larsen y Buss (2005) establecen diversas variables influyentes en la significación tales como: 1) el grado de correspondencia de los hallazgos obtenidos con relación a la persona que los denota; validez; 2) el grado en que una medida obtenida representa el nivel verdadero del rasgo a medir; confiabilidad y 3) generalizabilidad, es decir, el grado en que la medida conserva su validez y confiabilidad a lo largo de diversas situaciones (diversos grupos de personas, culturas, estatus socioeconómicos, etc.) y condiciones (lugar, materiales, ambiente, etc.). De manera general, se establece que los conceptos de validez y confiabilidad pueden incluirse en la generalizabilidad, haciendo alusión a la pregunta de ¿Cuáles son los contextos a lo largo de los cuales se pueden generalizar las puntuaciones de los instrumentos?

En México, la evolución constante de la tecnología y demanda social han impulsado el crecimiento de la psicología con respecto a estrategias de aplicación y evaluación que conlleven medios cibernéticos; desde su valoración como medio clave del método seleccionado o como factor influyente en el estudio a realizar, siendo de esa manera el ámbito Organizacional el más afectado por esta demanda digital-laboral (Aamodt, 2010).

Se ha observado que existe la necesidad de emplear pruebas psicológicas a través de plataformas en línea para la selección y reclutamiento de personal en masa o individual; capacitación en línea, evaluación del desempeño mediante pruebas virtuales, etc. (Molinar, Escoto, García y Bautista, 2012; Morales-Ramírez, Escoto, García-Lozano, Molinar-Solís e Hidalgo-Cortés, 2012; Olea, Abad y Barrada, 2010; Sierra–Matamoros, Valdelamar–Jiménez, Hernández–Tamayo y Sarmiento–García, 2007; Spector, 2002). No obstante, debido a esta necesidad y a la rapidez con la que se intenta dar solución a esta demanda, no se conoce a ciencia cierta qué tanto se están modificando los componentes esenciales que constituyen a las pruebas psicológicas estandarizadas; confiabilidad, validez, grado de significancia de respuestas y factores que intervienen en la resolución de una prueba (errores de medición).

Andrade, Navarro y Yock (1999, citados en González, Aragón y Silva, 2000; Zúñiga-Brenes, 2007) exponen que éstos factores suelen generar graves problemas en la captura, análisis e interpretación de los datos si no son controlados de manera adecuada; ya que son complicados de identificar y, por lo tanto, de corregir; principalmente aquellos que van en función a la experiencia individual del participante. Zúñiga-Brenes y Montero-Rojas (2007) sostienen que dentro la teoría de la generalizabilidad se establece este factor como que una de las principales variables que afectan en la resolución de una prueba psicométrica. Ya que la manera en que el evaluado se percibe en una situación de evaluación; antes, durante y después de ésta puede invalidar la evaluación que se está suscitando.

Catalán y González (2010), exponen en su investigación *"Actitud hacia la Evaluación del Desempeño Docente y su Relación con la Autoevaluación del Propio Desempeño, en Profesores Básicos de Copiapó, La Serena y Coquimbo",* la importancia que tiene la experiencia y actitud que los participantes presentan para con la evaluación sobre los puntajes obtenidos. En este estudio, ellos encontraron que *la* actitud y experiencia que los participantes tengan para con el test está directamente relacionada con las puntuaciones obtenidas a través de estos instrumentos; mismos que sostienen que a mayor experiencia positiva mayor puntuación favorable.

Por otro lado, existen estudios que abarcan el uso de tecnologías de la información como medios para ejecutar pruebas psicométricas de manera masiva y se asegura que

estas herramientas pueden impulsar de manera exponencial el uso de pruebas psicométricas digitales (Campos, Quero, Bretón, Riera, Mira, Tortella y Botella, 2015, Aguilar-Espinoza, 2013, González y Hernández, 2013, Bruning, Schraw, Norby, 2012). No obstante, en ninguna de ellas hace alusión a la experiencia individual de los participantes u otros factores actitudinales y aptitudinales que pueden generar afectaciones importantes en el estudio e incluso invalidar la prueba psicométrica que se está empleando.

Con base en esto, surge la necesidad de conocer el tipo de experiencia que los participantes denotan después de haber concluido una prueba psicométrica sistematizada y cómo es que esta afecta a las puntuaciones obtenidas en el instrumento. Es, por tanto, que el objetivo de este estudio es conocer la experiencia de los participantes con respecto a la resolución de una prueba psicométrica sistematizada.

Con el fin de lograr el objetivo se tomó como base el test de inteligencia “Factor G” de Cattell ya que el número de ítems que lo componen es reducido en comparación con otros instrumentos (como bien es el caso de MMPI-2) y porque fue baremado en población mexicana en el año 2000 por González, Aragón y Silva; lo cual permite obtener puntajes válidos y confiables con respecto al constructo que se pretende medir.

1. ANTECEDENTES DE LA PSICOMETRÍA

1.1. ¿Qué es la psicometría?

Antes de comenzar concretamente con el tema de psicometría es necesario mencionar que en muchas ocasiones el término "Psicometría" tiende a confundirse con "La evaluación psicológica", razón por la cual es necesario hacer una diferenciación antes de dar pie al tema como tal (Barbero, Vila y Holgado, 2010).

En relación a la última se puede decir que "aquello que llamamos Evaluación Psicológica no es más ni menos que un proceso de toma de decisiones" (Cronbach y Gleser, 1957, citados en Fernández, Noelia y Antonio, 2009, pp. 8) ya que el propósito del último es la recomendación de un camino de acción determinado en virtud de los objetivos perseguidos por la evaluación. De tal suerte que el psicólogo funge únicamente como un evaluador y no como aquel que toma la decisión de llevar a cabo las acciones o medidas recomendadas por este.

Es así como la evaluación psicológica se muestra como una tarea psicológica aplicada dirigida a la solución de problemas personales, instituciones, grupos, comunidades, sociales o ambientales, siendo así el uso de un modelo teórico como parte esencial para comprender o analizar el fenómeno concreto que es objeto de nuestra atención; ello conlleva a que entre las actividades implicadas en la evaluación psicológica se encuentra la categorización, la comparación, el análisis y la contratación de datos referidos a atributos del sujeto y/o de la situación o interacción que se está analizando (Anguera, 1995, Casullo, 1996, Fernández, 1993, Forns y Santacana, 1993 y Silva, 1990, citados en Fernández, Noelia y Antonio, 2009). Y es justamente este marco teórico el cual, en efecto, condicionará cada paso que realicemos en cada evaluación, es decir, los conceptos serán diferentes a los de otras disciplinas, las herramientas que se deseen emplear podrán modificarse con el fin de lograr aprehender el fenómeno, la interpretación de los datos será concretamente fiable en términos de la materia, etc.

Para combatir esto, dentro de la Evaluación Psicológica se han elaborado y concretizado diversas herramientas o instrumentos objetivos (a partir de un modelo teórico que empate con el fenómeno que se pretende medir), que resultan de suma importancia científica a causa de su gran elaboración estadística. Y es con ello con los cuales se puede recabar información necesaria o solicitada para cumplir el objetivo del proceso de evaluación y con ello llegar a una toma de decisión pertinente. Es así que todos los instrumentos, test, baterías, cuestionarios, escalas, etc., que se elaboraron a partir de un marco teórico psicológico adoptan el nombre de pruebas psicométricas; entendiéndose de esta manera a una prueba como a un desarrollo tecnológico derivado de los modelos teóricos de corte psicométrico (Fernández, Noelia y Antonio, 2009).

De acuerdo con Barbero, Vila y Holgado (2010) en un inicio se contemplaba a la Psicometría con base a su raíz etimológica griega que se conformaba por "Psykhe" y "Metrum", que literalmente significaba "Medida de la Psykhe"; sin embargo, ello resultaba sumamente subjetivo, de tal suerte que no fue sino hasta el estudio exhaustivo de las definiciones de varios expertos en la materia como Cerdá, (1970), Cliff, (1979), García-Cueto, (1993), Macià, (1982), quienes fueron retomados por Barbero et al. (2010) para formular una definición un tanto más clara sobre la Psicometría, por lo cual la definieron de la siguiente manera:

> "La psicometría es una disciplina metodológica, dentro del área de la Psicología, cuya tarea fundamental es la medición o cuantificación de las variables psicológicas con todas las implicaciones que ello conlleva, tanto teóricas (posibilidades y criterios de medición) como prácticas (cómo se mide y con qué se mide)" (Barbero, Vila y Holgado:5).

Partiendo de la definición anterior, se puede concluir que la Psicometría es una disciplina encargada de justificar y legitimar la medición psicológica, misma que consistirá en desarrollar estrategias o paradigmas que permitan aprehender los fenómenos que se pretenden estudiar, por tal motivo se suscita la transformación de los hecho en datos cuantificables, mismo que a su vez genera la necesidad de analizar sobre el nivel o el grado en que dichos datos captan o representan realmente eso que dicen o aseguran medir. Es por ello que la psicometría no sólo nos permite analizar los eventos generados

en el exterior, sino que también da pie a establecer condiciones en las cuales son permitidos o pertinentes llevar a cabo las mediciones estadísticas.

1.2 ¿Para qué sirve la psicometría?

A pesar de que la psicometría no tiene un campo de aplicación especificó como sucede con otras disciplinas, su aplicación abarca todas las áreas de la Psicología tales como la personalidad, inteligencia, procesos cognitivos, procesos neuropsicológicos, actitudes, aptitudes, etc., y que, a su vez, nos permite contrastar, fundamentar o elaborar nuevas teorías. Ello a causa de que a ésta disciplina le incumbe todo lo relacionado con la medición de variables psicológicas, de tal manera que ello justifica su implementación si se tiene en cuenta que en esta profesión al igual que en las demás ciencias empíricas, el objetivo final es la descripción, explicación y predicción de los fenómenos de interés.

Ahora bien, se estipula que la base de la psicometría es la medición, pero, ¿A qué nos referimos con este concepto? En la vida diaria la palabra medición tiene un significado claro y conciso, se conoce que para poder dar cuenta de algo es necesario llevar a cabo un proceso valorativo del mismo que se hace a través de diversas herramientas. Estas herramientas a su vez proporcionarán datos fijos y precisos en forma de puntajes que denotará, por ejemplo, centímetros, gramos, segundos, etc. En este caso, se usan instrumentos físicos que a *grosso modo* no presentan mayores problemas en la praxis y en el proceso de interpretación.

No obstante, ¿Qué ocurre con las mediciones psicológicas realizadas por las pruebas psicométricas? Aquí la situación es totalmente diferente, las variables que se pretenden medir no son del todo concretas y observables en primera instancia. Las variables se definen como aquellas propiedades o características que poseen diferentes individuos en cantidades distintas (Mugnusson, 1977 y Mendenhall, 2017).

Partiendo de esta idea, la definición de medición es clara en tanto que se refiere a la asignación de números a las propiedades o características de los objetos de acuerdo con las reglas dadas, cuya validez puede probarse empíricamente y que permita con ello

representar las cantidades en atributos (Mugnusson, 1977, Nunnally, 1987, Barbero et al, 2010). Esto quiere decir que partimos de la idea de que los atributos son características específicas y particulares de los objetos, razón por la cual se estipula que la medición psicométrica realizada no es al objeto *per se*, sino a los rasgos que lo identifican como ese objeto determinado.

Es por ello que las reglas son determinantes a la hora de obtener dichos atributos, en este sentido, Nunnally (1987) sostiene que el término de "reglas" hace alusión al procedimiento explícito para asignar números a los diversos factores en función del carácter del mismo. Aquí los números conllevan un papel fundamental, puesto que al ser medidas estandarizadas nos proporcionan objetividad, posibilidad de cuantificación, comunicación y economía (en relación al tiempo y dinero), lo cual a su vez impulsa un crecimiento científico de la profesión con base a una medición eficiente.

Ahora bien, al establecer reglas es necesario destacar que estas no son ambiguas. No obstante, pueden desarrollarse a partir de un modelo deductivo a partir de un gran número de experiencias previa y que tienen como objetivo dar solución a la pregunta de qué tanto sirve la medida para explicar el fenómeno (Nunnally, 1987).

En este sentido, las pruebas psicométricas se usan en función a la utilidad que éstas aportan a la explicación, descripción y/o correlación de una situación, constructo o variable; mismos que dan pauta al investigador para definir hasta qué punto la medida que realiza tiene relación objetiva con lo que se pretende medir. Para lograr esto generalmente se emplean procesos estadísticos *inferenciales* y *descriptivos* que permiten al investigador realizar afirmaciones probabilísticas que se relacionan con los valores observados en la muestra y llevar a cabo procesos matemáticos que faciliten el análisis de las medidas de tendencia central y la dispersión de los datos empíricos. Mismos que guiarán sin duda al investigador a una pertinente y objetiva interpretación de los datos a partir de la valoración cuantitativa obtenida a través de estos procesos estadísticos (Nunnally, 1987).

Por otro lado, llevar a cabo este tipo de procesos estadísticos no es una tarea fácil para el investigador y en numerosas ocasiones suceden errores de interpretación a causa de un mal uso de los procesos estadísticos inferenciales y descriptivos. En lo que va a este

tipo de adversidades Barbero et al, (2010) plantea que los problemas de interpretación hacen alusión a la legitimidad de emplear clases particulares de procedimientos matemáticos en la medición de atributos psicológicos para el control y uso de los datos; los cuales son conocidos como *formas fundamentales referidas a la norma* y *referidas al criterio*, mismas que a su vez se subdividen en escalas de medición: Ordinal y Nominal (referidas a la norma) e Intervalo y de Razón (referidas al criterio).

Los datos referidos a la norma consisten en comparar resultados obtenidos por un sujeto en contraposición con los obtenidos por otro grupo de sujetos que conforman el grupo normativo y que pertenecen a una misma población. Por otra parte, los procesos de control de datos referidos al criterio hacen referente a la relación de una razón previamente establecida; por ejemplo, los resultados obtenidos se comparan con esta razón o criterio (punto crítico) y la superación o no del mismo es lo que va a dar significado a las puntuaciones obtenidas (Barbero, Vila y Holgado, 2010). Mendenhall, (2017) define a las escalas de medición de la siguiente manera:

1.2.1. Escala Nominal

El primer tipo de escala es conocido como Nominal, y éste se encarga únicamente de nombrar a los diferentes factores, objetos, variables, atributos, etc., que se produzcan a partir del proceso de medición y valoración. Este sería el caso de usar el 0 para sexo femenino y el 1 para el masculino (o viceversa) o asignar letras para reconocer a los distintos individuos de una población evaluada.

1.2.2. Escala Ordinal

Los datos proporcionados a través de la medición solamente nos permitirán ordenar los objetos con respecto al rasgo que se midió. Y ya que el orden de los objetos es la única información transmitida por los números, éstos pueden reemplazarse por diversos signos o símbolos que permitan secuenciar los datos (Mugnusson, 1977 y Mendenhall, 2017).

Mendenhall (2017), sostiene que cuando se realiza una medición psicológica, generalmente no se llega más allá del nivel ordinal si no se tienen suposiciones o teorías sobre el fenómeno. Sin embargo, resulta fundamental el orden de los datos en toda

investigación de corte cuantitativo puesto que nos permite conocer la posición de cada individuo en el continuo que expresa la magnitud del rasgo que posee y, con ello, la posibilidad de describir la posición de un individuo solamente como más grande que, igual a, o menor que, de la posición de otro individuo. Es decir, la regla de correspondencia aquí nos permite asignar los valores numéricos a una propiedad del objeto de estudio, de tal forma que reflejen niveles crecientes de esa propiedad sin que haya un compromiso de que las distancias en esa propiedad sean iguales.

1.2.3. Escala de Intervalos

En este nivel de medición se lleva a cabo el uso de los procedimientos de las escalas anteriores, es decir, los datos se nombran (nominal) y se ordenan (ordinal) en función a las necesidades requeridas en la investigación, sin embargo, aquí también se establece una distancia entre número y número, las reglas que imperan en este nivel es que la distancia que exista entre un dato y otro debe ser el mismo. Por ejemplo, un incremento de 5ºC es igual, ya sea cuando se pasa de 0 a 5ºC o cuando se pasa de 10 a 15ºC. Y, por último, que a diferencia de las demás escalas aquí aparece por primera vez el 0, sin embargo, el cero no es absoluto, sino que funge como una referencia de ubicación entre intervalos o números.

1.2.4. Escala de Razón

En este último nivel de escala, se usan las propiedades anteriores, pero, además se tiene un cero absoluto que refleja la ausencia de la cualidad. Por ejemplo, en el caso anterior de la temperatura las escalas hacen referencias a un cero que es arbitrario y no refleja la ausencia de valor, sino el punto en el cual el hielo se derrite (o el agua se congela) de acuerdo a las leyes que sustentan la teoría de los grados Celsius. Sin embargo, en lo que va del teorema de los grados Kelvin, aquí el cero absoluto hace referencia a la ausencia total del movimiento molecular y, por lo tanto, la ausencia de temperatura.

Es de esta forma que los diferentes tipos de escalas usan ciertas propiedades de los sistemas numéricos para generar un tipo de medidas que reflejen ciertas propiedades de la dimensión que se pretende reflejar con esas medidas.

Mugnusson (1977) y Mendenhall (2017), establecen que algunas de las limitantes de las estadísticas a partir de los niveles de medición es que, no cualquier investigación puede llevarlas a cabo y que, al mismo tiempo, el tipo de medida determinará el tipo de estadístico a emplear. Por ejemplo, en el caso de las mediciones de intervalo y de razón se utilizan procesos más complejos y elaborados que las demás mediciones; éstos comprenden a las mediciones de corte paramétrico. Mientras que las mediciones no tan complicadas como lo son las nominales y ordinales son conocidas como mediciones no paramétricas y, algunos estadísticos pertenecientes a este corte son, por ejemplo, Kolmogorov-Smirnov o la U de Mann-Whitney que se usan para las mediciones de propiedades referidas al orden. Mientras que, por otro lado, se encuentra la Ji cuadrada, la cual se emplea para determinar las probabilidades de clases de eventos.

Con relación a los métodos que existen para evaluar la calidad métrica de las medidas obtenidas Muñiz (1992 citado en Barbero, Vila y Holgado, 2010) indica que la única manera de hacerlo es a través de técnicas estadísticas encuadradas bajo las denominaciones de fiabilidad y validez; los cuales tienen por objetivo fundamental estimar los errores aleatorios que conlleva toda medición (confiabilidad de las medidas) y, segundo, garantizar que la misma no es algo inútil sino que sirve para explicar y predecir los fenómenos de interés (validez de las medidas) (Barbero, Vila y Holgado, 2010).

1.3. ¿Cuándo aparecieron por primera vez las pruebas psicométricas?

Tomando como base el hecho de que la psicometría compone a la evaluación psicológica, ahora es necesario abordar el momento en el cual la evaluación psicológica comenzó a establecerse como disciplina y, cómo es que a través de ella y las diversas demandas sociales se establece la necesidad de una evaluación más rápida y eficiente para las diversas poblaciones pertenecientes a una sociedad específica; siendo ésta la principal razón por la cual nacen de los test psicométricos.

Ahora bien, al hablar de la constitución de la evaluación psicológica como disciplina científica hemos de citar a Francis Galton, McKeen Cattell y Alfred Binet, quienes fueron los pioneros en la construcción de pruebas psicométricas del siglo XIX, con ellos, se establecen las bases conceptuales, metodológicas y tecnológicas del psicodiagnóstico (Terán, 2017).

Sin embargo, los factores que eran medidos en ese entonces con ayuda de las pruebas psicológicas diferían a las que se tienen hoy en día. En aquella época existía el auge de los estudios realizados por Darwin de las especies que se encontraban en las Islas de los Galápagos. Mismos estudios impulsaron a su primo Galton a formular medios o herramientas que permitieran conocer quiénes eran los humanos más capacitados para sobrevivir y, con ello, crear una sociedad conformada por los humanos más dotados del mundo. Las primeras herramientas se basaron en la evaluación de las diferencias entre personas y la medición de éstas, lo cual trajo como consecuencia el surgimiento de los instrumentos de evaluación psicológica (Aragón, 2015, 2010, 2004; González, 2007).

Ello fue la razón por la cual Galton fundó en Londres en 1884 un Laboratorio Antropométrico donde realizó arduas mediciones a las personas sobre su estatura, peso, capacidad auditiva, agudeza visual, capacidad sensorial discriminativa, percepciones y capacidades motoras. De esta forma introdujo las bases de la evaluación cuantitativa de las diferencias humanas con base a la sistematización de la recogida de datos y su tratamiento estadístico con lo cual dio inicio al estudio psicológico de las diferencias individuales frente a la psicología experimental a finales del siglo XIX. Es por todo esto que a Galton se le considera actualmente el fundador de la Psicología diferencial (Muñiz, 2010).

Aragón (2015), Muñiz (2010) y González (2007) sostienen que Galton contaba con muchas personas que lo apoyaban en sus estudios sobre las diferencias individuales en su laboratorio en Londres, y uno de sus más grandes estudiantes, Cattell, psicólogo estadounidense, 1861-1934, fue otro de los grandes personajes que impulsó la evaluación psicológica.

Terán (2015), Muñiz (2010) y González (2007) sostienen que fue él quien introdujo el concepto de test mental en 1890; además publicó varias pruebas sobre

ejecuciones específicas de los sujetos a nivel sensorial, perceptivo y motor, destacándose en el estudio diferencial de los tiempos de reacción; rechazó la introspección como método de estudio propuesta por Wilhelm Wundt a finales del siglo XIX y se enfocó en la necesidad de que las medidas obtenidas en los test fueran objetivas, con ese fin planteó el uso de baterías de pruebas para la evaluación psicológica, en el año 1896, introduciendo de este modo el concepto de batería de pruebas, mismas que se definían como una serie de test administrados anualmente a estudiantes universitarios con el propósito de determinar su nivel intelectual; estas pruebas incluían mediciones de fuerza muscular, agudeza auditiva y visual, sensibilidad al dolor, memoria, velocidad de reacción y otras más.

Otro gran personaje dentro de la evaluación psicológica y que llevó a cabo pruebas o test psicométricos para evaluar las diferencias individuales, de acuerdo con González (2007) y Muñiz (2010), fue Alfred Binet, 1857-1911, en Francia. Los autores mencionan que él marcó un avance cualitativo importante en el estudio de las diferencias individuales ya que propuso un nuevo enfoque en la evaluación psicológica. Su objetivo no fueron las diferencias de las funciones sensoriales, perceptivas y motoras planteadas por Galton y Cattell; su interés en las diferencias individuales se dirigió a la evaluación de las funciones psíquicas superiores. Para lograr ese objetivo, Muñiz (2010) menciona que él planteó el método de los test mentales, y preocupado por la objetividad de estos instrumentos, expuso lo siguiente:

- Las pruebas deben ser sencillas.
- Su aplicación debe tener una poca inversión de tiempo.
- Debían ser independientes del examinador y
- Los resultados obtenidos pudieran ser contrastados por otros observadores.

Por otra parte, González (2007) menciona que Alfred Binet consideró tres métodos diferentes en el trabajo diferencial del niño normal y el retrasado mental: 1) el

examen médico, 2) el examen escolar (realizado por el profesor) y 3) el diagnóstico psicológico que evalúa los procesos mentales superiores del sujeto mediante su ejecución en una prueba, dando lugar así, junto a su colega Theodore Simon, al primer test de inteligencia, en 1905. Esta escala de inteligencia fue adaptada en 1960 y en 1973, y a partir de ese año el instrumento adoptó el nombre de Terman Merrill.

Comenzaba así una expansión creciente en el uso y creación de test de todo tipo. González (2007), expuso que fue de esta manera en la que la técnica del análisis factorial permitió un gran avance en la construcción y análisis de los test, ya que ellos dieron pie a la aparición de las baterías de test cuyo representante más genuino serían las Aptitudes Mentales Primarias (PMA) de Thurstone (Thurstone, 1938; Thurstone y Thurstone, 1941 citados en Moreno, 2007).

Martínez, Hernández y Hernández (2014), sostienen que existe una división en la manera de concebir y definir a la inteligencia; tomando como base a los distintos factores o dimensiones que la componen; permitiendo de esta manera la aparición de dos grandes líneas de estructuración de las dimensiones cognitivas: la escuela inglesa y la escuela americana. En la primera se da más importancia a un factor central de inteligencia general, que vendrían dos amplias dimensiones: la verbal-educativa y la mecánico-espacial. El enfoque americano asume una serie de dimensiones no jerarquizadas que compondrían el perfil cognitivo, por ejemplo, en el caso del PMA serían: la comprensión verbal, la fluidez verbal, aptitud numérica, aptitud espacial, memoria, rapidez perceptiva y razonamiento general. Ambos enfoques son compatibles, y tienen mucho que ver con la tecnología estadística utilizada, sobre todo el análisis factorial (Terán, 2017 y Matesanz, 1998).

Toda esta línea de investigaciones psicométricas sobre la inteligencia culmina en la obra magna de Carroll (1993), quien sintetiza los grandes avances alcanzados con respecto a este tema. En España trabajos como los de Juan-Espinosa (1997), Colom (1995) o Andrés-Pueyo (1996) recogen y analizan de forma brillante este campo de trabajo (Muñiz, 2010).

De manera general, estos tres autores son los iniciadores de la constitución de nuestra disciplina y comparten ese mérito con otros científicos, como por ejemplo los primeros psicólogos matemáticos Pearson y Spearman, que en la última década del siglo

XIX dan lugar a importantes técnicas estadísticas que son la base matemática necesaria para los estudios de grupo propios de la Psicología diferencial, a través de los cuales construyeron los primeros test psicométricos y que, no obstante, se siguen empleando para la construcción de nuevas baterías de pruebas psicológicas (Martínez et al, 2014 y Muñiz, 2010).

1.4. Modelos de creación de las pruebas psicométricas

A la construcción y análisis de los test subyacen teorías que guían su creación, que condicionan y tiñen a éstos según los avances teóricos y estadísticos de cada momento. Hay dos grandes enfoques que resultan vitales a la hora de construir y analizar los test, ellos son la Teoría Clásica de los Test y el enfoque de la Teoría de Respuesta a los Ítems (Mendenhall, Beaver y Beaver, 2017). Ahora bien, no se trata aquí de llevar a cabo una profundización exhaustiva de estas teorías, sino tan sólo subrayar los aspectos claves sobre los mismos para que así los usuarios de los test tengan una idea más clara y, con ello, logren comprender en profundidad el alcance de las propiedades psicométricas de los test que están utilizando.

De acuerdo con Muñiz (2010), el psicólogo como cualquier otro profesional de otro campo, tiene que asegurarse de que el instrumento que utiliza mide con precisión; con poco error. No es difícil estar de acuerdo en esto, pero el problema es que cuando un psicólogo aplica un test a una persona (o varias), lo que obtiene son las puntuaciones empíricas. Sin embargo, no sabemos si esas puntuaciones empíricas obtenidas corresponden o no con las puntuaciones que verdaderamente pertenecen a esa persona en la prueba, y es allí donde entra el papel del psicólogo para garantizar que las puntuaciones obtenidas tienen un margen de error mínimo. En otras palabras, el error está mezclado con la verdadera puntuación, como la sal en el agua del mar y para separarlos necesitamos llevar a cabo algunos procesos y ahí es donde entra la teoría clásica de los test compuesta por el modelo Lineal Clásico de Spearman-Brown.

1.4.1 Modelo Lineal Clásico de B. Spearman

De acuerdo con Terán (2017), la psicometría es una rama relativamente muy joven de la psicología y, el núcleo que consolida a ésta rama subyace en las investigaciones realizadas por Spearman en los años 1904 a 1913 (Muñiz, 1992 citado en Aragón, 2004), cuyo objetivo era encontrar un modelo estadístico que fundamentase las puntuaciones de los test y permitiera la estimación de los errores de medición asociados a todo proceso de medición. Actualmente conocido como el modelo lineal clásico de Spearman; mismo que se expresa como: X = V + e, en donde "X" es la puntuación empírica de un sujeto, "V" es la puntuación verdadera y "e" es el término de error.

Aragón (2004), sostiene este primer supuesto es la principal aportación de Spearman, mismo que estipula que la puntuación de un sujeto en una prueba es igual a la puntuación verdadera (que nunca se puede conocer), más el error de medición; lo que significa que mientras más confiable sea la prueba y por tanto menor sea el error la puntuación empírica se acercará más a la puntuación verdadera. Reconocer que toda medida psicológica, al igual que las medidas de la ciencia contienen un término de error, fue un gran avance para la evaluación psicológica, ya que entonces se empieza a considerar que no basta el juicio humano para determinar los atributos psicológicos, y que los instrumentos de evaluación psicológica deben intentar, al igual que los instrumentos de las ciencias físicas, medir con precisión.

Por otra parte, Muñiz (2010) y Fernández, Noelia y Antonio (2009), exponen que los otros dos supuestos son igual de relevantes para la psicología al igual que el primero, ellos describen el segundo supuesto de Spearman como aquel que asume que no existe relación entre la cuantía de las puntuaciones verdaderas de las personas y el tamaño de los errores que afectan a esas puntuaciones. En otras palabras, que el valor de la puntuación verdadera de una persona no tiene nada que ver con el error que afecta esa puntuación, es decir, puede haber puntuaciones verdaderas altas con errores bajos, o altos, no hay conexión entre el tamaño de la puntuación verdadera y el tamaño de los errores. De nuevo se trata de un supuesto en principio razonable, que formalmente puede expresarse así: r (v e) = 0.

El tercer supuesto establece que los errores de medida de las personas en un test no están relacionados con los errores de medida en otro test distinto. Es decir, no hay ninguna razón para pensar que los errores cometidos en una ocasión vayan a covariar sistemáticamente con los cometidos en otra ocasión. Formalmente este supuesto puede expresarse así: r (ej., ek) =0 (Muñiz, 2010 y Fernández et al, 2009).

Además del modelo y de estos tres supuestos, se formula una definición de lo que son Test Paralelos, entendiendo por ello aquellos test que miden lo mismo exactamente, pero con distintos ítems. Las puntuaciones verdaderas de las personas en los test paralelos serían las mismas, y también serían iguales las varianzas de los errores de medida. Pues bien, el modelo lineal, junto con los tres supuestos enunciados, y la definición de test paralelos propuesta, constituyen el cogollo central de la Teoría Clásica de los Test (Magnusson, 1977).

Entonces, para poder considerar a las pruebas psicométricas adecuados y científicas, éstas deben contar con dos requisitos indispensables: la confiabilidad y la validez. De modo que cuando los psicólogos manipulen sus coeficientes de fiabilidad y validez para indicar a sus clientes o usuarios en general que los test que utilizan son precisos han de saber que esa estimación de la fiabilidad se puede hacer gracias a este sencillo modelo y a los supuestos planteados hace ya más de cien años (Prieto y Delgado, 2010 y Muñiz, 2010).

2. PRINCIPIOS DE LOS TESTS PSICOMÉTRICOS

De manera general se establece que para garantizar una medición pertinente, objetiva y con un margen de error mínimo se debe tener muy claro los diversos factores que aseguran que una prueba sea eficiente o no. Mislevy *et al* (2003, citado en Aragón, 2015), establece que los factores más importantes de una prueba son los coeficientes de validez, confiabilidad que el instrumento posea, los errores de mediciones que pueden acontecer en el proceso de evaluación y los esfuerzos realizados por reducir al máximo éstas variables interventoras.

2.1. Confiabilidad

A diferencia de otros factores, la fiabilidad no se establece en función de "hay" o "no hay", sino que se parte del supuesto teórico de que, en efecto, "... todas las pruebas psicológicas cuentan con fiabilidad, la cuestión aquí es con qué de confiabilidad cuenta una prueba psicométrica" (Muñiz, 2003 citado en Aragón 2015, pp. 17). Es por ello que éste es un continuo que abarca desde la consistencia mínima de una medición hasta casi llegar a la perfección de repetición de los resultados.

Este principio tiene que ver con los errores cometidos en el proceso de medición, por lo que responde al problema de hasta qué punto las cantidades observadas reflejan con precisión la puntuación verdadera de la persona. Se establecía en la teoría clásica de los test que una puntuación obtenida empíricamente por un sujeto es igual a la puntuación verdadera (la cual nunca se puede conocer en su totalidad) más el error de medición; lo cual indica que, mientras más confiable sea la prueba entonces menor será el error cometido en la evaluación (Aragón, 2004, 2015 y Martínez et al, 2014), es decir, de manera general se puede definir a la fiabilidad como "... la precisión con que un instrumento mide un objeto; en términos estrictos, la confiabilidad sería la ausencia de errores de medición" (Aragón, 2015, pp. 43).

En psicología ello resulta sumamente complicado a causa de que al medir los atributos psicológicos tenemos que hacerlo a partir de sus rasgos específicos, ya que los

fenómenos psíquicos no se pueden medir de manera directa. Es por ello que para minimizar los errores se toman en cuentas dos formas adyacentes que conforman a la fiabilidad de un instrumento: Confiabilidad de estabilidad temporal y Confiabilidad de consistencia interna.

2.1.1. Confiabilidad de Estabilidad Temporal

La estabilidad temporal indica el grado en que las puntuaciones de un test quedan afectadas por las fluctuaciones diarias que se producen en el sujeto o en el ambiente en que se aplica el test (Nunnally, 1980). La estabilidad temporal de un test depende parcialmente de la longitud del intervalo sobre el que se mide (González, 2007), por ejemplo, si disponemos de las puntuaciones de N personas en un test y, después de transcurrido un tiempo, volvemos a medir a las mismas personas con el mismo test, cabe suponer que, siendo el test altamente fiable, deberíamos obtener una correlación de Pearson elevada entre ambas mediciones. Dicha correlación entre la evaluación test y la evaluación retest (rxx) se denomina coeficiente de fiabilidad test-retest, e indicará tanta mayor estabilidad temporal de la prueba cuanto más cercano a uno esté (Muñiz, 1992, citado en Aragón, 2015; Abad et al, 2006).

Esta forma de actuar se desprende directamente del modelo lineal clásico, el cual se define a la fiabilidad como la correlación entre las puntuaciones empíricas en dos formas paralelas, ya que no existe mayor grado de paralelismo entre dos test que cuando en realidad es uno mismo aplicado dos veces. Este coeficiente se obtiene, sobre todo, en pruebas cuyo objetivo de medida es un rasgo estable (pruebas de inteligencia general, aptitudes, rasgos de personalidad, etc.) dado que, de lo contrario, no se podría discernir entre la inestabilidad del rasgo y la inestabilidad del instrumento de medición (Abad et al, 2006).

El valor de la confiabilidad indica el porcentaje de varianza en las puntuaciones obtenidas que es explicado por la variabilidad en las puntuaciones verdaderas y en qué medida se explica por efectos aleatorios. Por ejemplo, un valor de confiabilidad de 0.85 indicaría que la puntuación del sujeto se explica en un 85% por puntuaciones verdaderas recabadas y un 15% por causas no determinadas (Aragón, 2015).

Debe tenerse en cuenta, sin embargo, que cuanto mayor es el intervalo temporal que se deja entre ambas aplicaciones, mayor es la posibilidad de que las puntuaciones de los sujetos oscilan diferencialmente debido a factores de tipo madurativo y, por lo tanto, esto tiene un efecto concreto en el decremento de la correlación entre las puntuaciones del test y del retest (Nunnally, 1980).

2.1.2. Confiabilidad como consistencia interna

La precisión o fiabilidad de un test se puede entender también como el grado en que diferentes subconjuntos de ítems miden un rasgo o comportamiento específico; es decir, el grado en que covarían, correlacionan o son consistentes entre sí las diferentes partes del cuestionario. La consistencia interna se refiere a que los reactivos de un instrumento son congruentes unos con otros en la forma en que evalúan el mismo atributo psicológico. Por ejemplo, los sujetos tendrán un puntaje elevado en los reactivos que tienden a medir ese atributo y, al mismo tiempo, tendrán puntajes bajos en los reactivos que no miden ese atributo (Aragón, 2004; Fernández, Noelia; Antonio, 2009).

Lo más usual, de acuerdo con Fernández, et al (2009), es obtener la consistencia entre dos test que midan exactamente lo mismo y que sean paralelos, sin embargo, ello resulta sumamente complicado, razón por la cual se desarrolló dos métodos en los cuales es posible obtener un coeficiente de confiabilidad en relación a la consistencia interna del instrumento. El primero método para lograr esto es a partir del procedimiento conocido como "método de dos mitades". Este procedimiento nos permitirá conocer la confiabilidad de la consistencia interna del test.

Este procedimiento, de acuerdo con Barbero et al (2010), Abad et al (2006) y Magnusson (1977) consiste en dividir el test original en dos mitades equivalentes (normalmente una con los elementos pares y otra con los impares). Para cada sujeto se obtiene la puntuación directa en ambas mitades. Disponemos entonces de dos variables (P e I), cuya correlación de Pearson (rPI) indica su grado de relación. Si la mitad par e impar fueran entre sí formas paralelas, la correlación entre ambas sería una medida de la fiabilidad de cada una de ellas. Ahora bien, cuando hemos deducido la fórmula general de Spearman-Brown hemos visto que los test más largos (con más ítems) suelen ser más

fiables, por lo que rPI estará subestimando al coeficiente de fiabilidad del test total en la medida que P e I son variables extraídas de la mitad de ítems que tiene el test. Para superar este problema, y así obtener el coeficiente de fiabilidad del test completo, debemos aplicar la fórmula de Spearman, considerando que ahora que estamos trabajando con datos muestrales, y haciendo n = 2 ya que el test completo tiene el doble de ítems que cualquiera de sus mitades.

A partir de esta fórmula podemos comprobar que el coeficiente de fiabilidad, entendido como la expresión de la consistencia entre dos mitades, es mayor que la correlación de Pearson entre ambas mitades. La razón de dividir el test en la mitad par y la impar es garantizar su equivalencia. Los test de rendimiento óptimo suelen tener ítems ordenados en dificultad, de tal forma que se comienza a responder los ítems más fáciles hasta llegar a los situados al final del test, que son los más difíciles. Si realizamos la partición en dos mitades atendiendo a su disposición en la prueba (la primera mitad formada por los primeros n/2 ítems, la segunda por los n/2 ítems últimos) difícilmente podría cumplirse que ambas tuvieran la misma media (Prieto y Delgado, 2010).

Mendenhall et al (2017), postulan que si bien es cierto que en este método se establece que si k de los ítems del test son paralelos, es decir, que se correlacionan los reactivos de ambas mitades, entonces se buscará emplear la fórmula general de Spearman-Brown, o bien, si la puntuación de los reactivos de la prueba es dicotómica, la fórmula a emplear sería Kuder-Richardson; en ambos casos se parte del supuesto de que ambas mitades son equivalentes y es como si se aplicaran dos pruebas cortas y equivalentes. Dichos autores afirman que este proceso sólo resulta efectivo para la medición de los atributos intelectuales; ello vuelve dudoso la aplicación de este proceso en la medición de objetos no intelectuales o emocionales, lo cual dio pie a la elaboración de nuevos estadísticos que permitieran obtener la fiabilidad de un instrumento con el mismo procedimiento y que permitiera su aplicación a diversos atributos independientes de la inteligencia.

Es aquí donde nace el segundo método de obtención de coeficiente de confiabilidad denominado “método de consistencia interna de los reactivos individuales”. El principal exponente de este proceso fue Cronbach (1951, citado en Aragón, 2015, Martínez et al, 2014, Usabiaga, Castellano, Blanco-Villaseñor y Casamichana, 2013,

Gómez-Benito, Hidalgo y Guilera, 2010, Muñiz, 2010, Olea et al, 2010, Prieto y Delgado, 2010, Zúñiga-Brenes, 2007, Gempp, 2006, Abad, Garrido, Olea y Ponsoda, 2006, Díaz, Botanero, y Cobo, 2003) quien sostuvo que para no depender de una sola división se podría obtener la media de los coeficientes por mitades de todas las posibles formas de dividir la prueba en dos, lo que dio paso a la confiabilidad medida por una fórmula equivalente a la anterior; es decir, con el cual es posible obtener el mismo resultado que con la fórmula de Spearman-Brown o Kuder-Richardson; a éste fórmula se le conoce como coeficiente de confiabilidad α de Cronbach.

La ventaja que proporciona el emplear el coeficiente Alfa de Cronbach es que nos permite identificar los reactivos que se comportan de manera inconsistente y, por lo tanto, produce un valor más bajo de confiabilidad que el que se esperaría si se eliminara de la prueba (Magnusson, 1977). Sin embargo, es importante señalar que el coeficiente alfa de Cronbach no es un coeficiente de fiabilidad si, como ocurre en la práctica totalidad de los test, los ítems no son paralelos. De tal manera que, en este caso, sólo suele considerarse una "estimación por defecto" de la fiabilidad, lo que significa que es igual al coeficiente (si los ítems son paralelos) o menor a éste (cuando no son paralelos). Debe interpretarse como un indicador del grado de covariación entre los ítems, y es aconsejable complementarlo con otras técnicas estadísticas (Abad et al, 2006).

De esta forma podemos resumir que una prueba es confiable psicométricamente en tanto que su producción es consistente, esto es, en tanto que al ser aplicada en repetidas ocasiones se obtendrán puntuaciones iguales o muy similares a la puntuación original. O bien, si los ítems que componen al test son congruentes entre sí en relación a la forma en que miden el atributo propuestos por el instrumento. Sin embargo, ¿Qué es lo que ocurre con el significado de la medición que realizó el instrumento?, es decir, ¿Qué fue “eso” que se midió? Para dar respuesta a ello habremos de recurrir al concepto de validez; otro factor esencial en la construcción y aplicación objetiva de las pruebas psicométricas.

2.2. Validez

Es quizá el factor más importante de los principios de las pruebas psicométricas, y éste nos habla del grado en que el uso que pretendemos hacer de las puntuaciones de los test está justificado. Supone examinar la red de creencias y teorías sobre las que se asientan los datos y probar su fuerza y credibilidad por medio de diversas fuentes de evidencia (Nunnally, 1980). La validez es entendida como el grado en que un instrumento psicométrico mide realmente lo que se propone medir; es el grado de adecuación, significación y utilidad de las inferencias específicas que pueden derivarse a partir de las puntuaciones de los test, teniendo en cuenta que lo que se valida no es el instrumento, sino la interpretación de los datos obtenidos por medio de él (Martínez, 1996 citado en Aragón, 2004).

De manera general, se puede encontrar con que un instrumento sea confiable sin ser válida, Aragón (2015) y Prieto y Delgado (2010), proponen como ejemplo el utilizar una cinta métrica y medir en una jarra el nivel donde se encuentra el agua (p.ej., 10.5 cm); esta medida es confiable, pero no valida: si cambiamos el agua a otro recipiente la medida será diferente y no podemos generalizar lo medido. En otras palabras, no es válido medir la capacidad de un recipiente con un instrumento que mida la longitud de los objetos. Ahora bien, cabe resaltar que la validez no hace referencia al instrumento en sí, sino aquello que es válido en una prueba es la interpretación de los datos obtenidos por medio de un procedimiento específico de medición (Nunnally, 1980). En otras palabras, son las inferencias que podemos deducir de la ejecución de un sujeto en una prueba lo que es realmente válido (Muñiz, 2002; González, 2007; Martínez, 1996; Aragón y Silva, 2002 citados en Aragón 2004).

Básicamente, todos los procesos para determinar la validez de una prueba conciernen a las relaciones entre la ejecución y otros hechos observables de manera independiente acerca de las características de la conducta que se estudia. Es por ello que los primeros trabajos sobre éste principio distinguían entre un número de variedades que lo conforman. Mismas variedades fueron propuestas por los *Standards of the American Psychological Association* y se establecen en función al tipo de prueba y uso al que esté destinada la misma, ellos son conocidos como: validez de contenido, validez criterio

(concurrente o predictiva) y validez de constructo (Anastasi, 1988). Y éstos, a su vez, se les considera como diferentes formas de evidencias para un único tipo de validez.

2.2.1. Validez de Constructo

De manera general, se entiende como validez de constructo a la extensión en la cual la prueba dice medir un rasgo o característica específica teórica de un objeto, la cual, a su vez, está sustentada por la acumulación gradual de información proveniente de diversas fuentes que la respaldan, esto es, los constructos son aquellos que determinan qué conductas han de seleccionarse para su debida observación y medición en situaciones específicas de aplicación (Aragón, 2004 y González, 2007).

En realidad, al hablar de información que sustente el constructo hacemos referencia a cualquier dato que haga referencia a la naturaleza de este rasgo por más pequeño o insignificante que parezca. De ésta manera se intenta saber qué propiedades psicológicas o de otra índole pueden explicar la varianza de esas pruebas, es decir, explicar las diferencias individuales observadas en las puntuaciones del instrumento; así, no se trata sólo de validar la prueba, sino que se valida también la teoría sobre la cual se sustenta (Kirsch y Guthrie, 1980, citado en Aragón, 2004 y Martínez et al., 2010).

Nunnally (1987), establece tres aspectos principales para establecer las medidas de validación que debe tener una prueba psicológica: 1) especificación del dominio de las conductas observables; 2) determinar hasta qué punto todas o algunas de esas conductas se correlacionan entre sí y 3) precisar si una, algunas o todas las medidas de tales comportamientos actúan como si midieran el constructo. Para obtener el índice numérico de la validez de constructo se utiliza también la r de Pearson, que relaciona los valores obtenidos en la prueba con medidas que supone teóricamente se correlacionan con el constructo (como la inteligencia y las calificaciones académicas), o bien, con otra prueba que mida el mismo constructo y que ya esté validada.

Autores como Magnusson (1977) y Nunnally (1987) sostienen que la validez de constructo incluye la planificación y ejecución de determinados estudios de investigación orientados a comprobar empíricamente que un test mida realmente el constructo o rasgo que se plantea. Aunque los métodos usados son sin duda variados, así como las técnicas

estadísticas para analizar los datos, podemos encontrar un común denominador a todos ellos y, de acuerdo con Abad et al (2006), estos pueden englobarse en las siguientes fases:

1. Formular hipótesis relevantes (extraídas de deducciones teóricas o del sentido común) en las que aparezca el constructo que pretendemos evaluar con el test. En definitiva, una hipótesis de trabajo consiste en poner en relación dos o más variables. Pues bien, una de esas variables a ser el constructo que pretendemos medir con el test.

2. Efectuar en la práctica mediciones oportunas de las variables o constructos involucrados en las hipótesis. La medición del constructo de interés se realizará con la prueba diseñada a tal efecto, que es la que pretendemos validar.

3. Determinar si se verifican o no las hipótesis planteadas. En el caso de que así sea, queda confirmado mediante una investigación que el test mide el constructo de interés ya que, de lo contrario, no habría razones lógicas para que se cumplieran las hipótesis formuladas. Si las hipótesis no se confirman no significa en principio que el test no es válido, ya que puede ser debido a que las hipótesis no estaban planteadas de manera adecuada, lo cual exigiría una revisión de la teoría subyacente.

2.2.2. Validez de Criterio

Dentro de la validez de criterio se habla de validez concurrente y validez predictiva. Argibay, (2006), sostiene que la diferencia entre ambas formas de validez, radica en la temporalidad del criterio. Si las puntuaciones del test se utilizan para predecir alguna medida del criterio que se va a realizar a futuro, sería validez predictiva. Si por el contrario relacionamos las puntuaciones del test con alguna medida del criterio tomada en el mismo momento sería validez concurrente. Por ejemplo, si se aplicara el EPQ-A (Cuestionario de Personalidad de Eysenck para Adultos), y lo utilizan para predecir qué pacientes depresivos, pasado un año, tendrán una mejor respuesta al tratamiento, sería validez predictiva. Si por el contrario se aplica el EPQ-Ay el Inventario de Depresión de Beck

(como criterio), en forma simultánea, y relacionan los puntajes de ambos instrumentos entre sí, para analizar en qué medida el Neuroticísmo puede predisponer a la adquisición de conductas depresiva, sería validez concurrente.

Cuando se ha hecho un estudio de validez concurrente y se ha establecido que tal variable de un test correlaciona con tal criterio, estamos prediciendo esa correlación, y eso es poder predictivo. O sea, que la diferencia entre ambas formas de validez, no tiene que ver con si son predictivas o no (desde un punto de vista científico), sino más bien con el diseño que involucran en cuanto a su dimensión temporal. La validez concurrente implica un diseño transeccional o transversal, los cuales como bien mencionan Sampieri, Collado y Lucio (1997), "recolectan datos en un solo momento, en un tiempo único"; mientras que la validez predictiva implicaría un diseño prospectivo (Argibay, 2006).

La validez referida a la predicción (o criterio) es aquella que se emplea para estimar a futuro un comportamiento, la cual, una vez determinada por el investigador, se elabora una serie de reactivos que estarán correlacionados con él (González, 2007, Nunnally, 1973, Aragón, 2015, 2004; Barbero et al, 2010). Para lograr esto es necesario que exista evidencia de que hay una relación entre las puntuaciones del test y las del criterio, y sólo entonces se relacionan los puntajes obtenidos de la prueba con los puntajes de la variable de criterio, obteniendo de esta manera la validez de criterio.

Abad et al., (2006), Barbero et al., (2010) y Prieto y Delgado (2010), describen que la correlación entre las puntuaciones en el test (X) y en el criterio (Y) se le denomina coeficiente de validez, lo designamos como rxy e indicará el grado en el que el test sirve para pronosticar con precisión el rendimiento en el criterio. El coeficiente de validez es una correlación de Pearson y, por tanto, su interpretación más inmediata se fundamenta en el denominado coeficiente de determinación, que es simplemente el cuadrado de la correlación y que indica la proporción de varianza del criterio que podemos pronosticar con el test. Así, un test con un coeficiente de validez de 0.5 indicará que explica un 25 % de la variabilidad o diferencias individuales en el criterio, mientras que el 75 % restante se debe a variables diferentes al test. El coeficiente de determinación se puede expresar de la siguiente manera:

Cuando, tanto en contextos aplicados como investigadores, se desea predecir de la forma más precisa posible las puntuaciones en un determinado criterio, es común utilizar más un proceso estadístico predictivo. Razón por la cual tiende a emplearse las ecuaciones o técnicas estadísticas de Regresión Múltiple o lineal (Terán, 2017).

Por otro lado, la validez de Criterio también puede obtenerse a partir de la medición del grado en que una prueba psicométrica se correlaciona con otro test que mida el mismo constructo; a este tipo de validez se le denomina "Concurrente". El cual establece que una nueva prueba puede obtener su validez de criterio a partir de su puesta en escena con respecto a otro instrumento que cuenta con altos grados de validez con respecto a la variable que pretende medir (Castillo y Folino, 2009).

2.2.3. Validez de Contenido

La validez aquí comprende el grado en que unos conjuntos de reactivos representan adecuadamente un dominio de conductas de interés (Linehan, 1980, citado en Aragón, 2004). Dicha validez consiste en medir una muestra representativa de un contenido teórico de conocimientos o habilidades; principalmente en pruebas de rendimiento (pruebas de inteligencia, de aptitudes, etc.) y en pruebas de conocimientos (cuestionarios para evaluar el rendimiento en una materia escolar o en una especialidad temática concreta), y con ello determinar si las respuestas dadas, así como las condiciones bajo las cuales las conductas observadas representan a todos aquellos conjuntos de respuestas y condiciones en las que se desenvuelve, son susceptibles a la generalización.

El objetivo de la validez de contenido, tal y como lo plantea Aragón, (2004) y Prieto y Delgado (2010), consiste en demostrar que los ítems de una prueba son representativos de un universo y con ello asegurar un muestreo cuidadoso de un dominio de contenido relevante. Al respecto de ello, también postula que existen dos aspectos esenciales y complementarios de la validez que deben ser tomados en cuenta en la construcción de instrumentos: 1) el instrumento no debe incluir aspectos irrelevantes de la conducta de interés y 2) que el instrumento contenga todos los aspectos importantes que definen el dominio conductual.

En definitiva, la validez de contenido es un tema particular de muestreo: si deseamos realizar inferencias sobre el rendimiento de las personas en una población de contenidos determinada, el test debe incluir una muestra representativa de dichos contenidos.

El proceso de validación de contenido es eminentemente lógico, si bien el método de validación no es meramente empírico, sí puede llevarse a cabo bajo un procedimiento racional-lógico como lo es el uso de jueces expertos en el tema para valorar la congruencia entre los diversos ítems y los diversos objetivos. Existen procedimientos cuantitativos diversos para que cada experto valore el grado en que un ítem sirve para evaluar el objetivo al que corresponde (Kerlinger, 1975; Hoste, 1981; Abad et al, 2006). Éstos se describen a continuación:

- Especificar los diversos objetivos (áreas diferentes de contenidos) que se pretenden evaluar.
- Elaborar varios ítems para cada objetivo.
- Seleccionar una muestra de expertos en el contenido del test.
- Pedirles que, según su opinión, asignen cada ítem al objetivo que pretende medir.
- Seleccionar los ítems en los que los expertos manifiestan mayor acuerdo en sus clasificaciones.

Si bien el objetivo de este trabajo no consiste en la elaboración de un instrumento de medición psicológica, es necesario conocer los principios de confiabilidad y validez que sustentan la aplicación y control de las puntuaciones obtenidas de una población. Ello permitirá al investigador o evaluador seleccionar una prueba pertinente y eficaz que le garantice resultados objetivos y precisos sobre los atributos psicológicos pertenecientes a una población en particular.

2.3. Errores de medición

De acuerdo con Gómez-Benito et al (2010), Fernández et al, (2009) y Gempp (2006), existen muchas formas de cometer errores a lo largo de la evaluación psicológica, sin embargo, aquellos que competen específicamente a las pruebas psicométricas se les conoce como errores aleatorios y no sistemáticos, ello a causa de que sus efectos son imposibles de predecir. En lo que va del primero e establece que contribuyen a una forma aleatoria a la puntuación de los sujetos; tanto pueden perjudicarlos como beneficiarlos dependiendo de las condiciones. Con respecto a los errores no sistemáticos, se establece que son los más riesgosos a la hora de establecer la confiabilidad de un instrumento, ya que puede que la prueba mida de manera consistente algo diferente e incluso contrario al propósito planteado. Por ejemplo, cuando un reactivo no es claro o induce la respuesta al evaluado.

Es por eso que el conocimiento preciso y exhaustivo de los factores que determinan la cuantía del coeficiente de fiabilidad resulta esencial para ayudarnos en la tarea de diseñar pruebas adecuadas. Empero si no se tiene un conocimiento profundo sobre el atributo entonces ello podría guiarnos a un sesgo científico importante en la evaluación psicológica.

De acuerdo con Magnusson, (1977), un error de medición que persigue a cada evaluador o autor de pruebas psicométricas es la selección objetiva de los reactivos a emplear en el instrumento. Éste debe saber con exactitud cuáles son los objetivos que se pretenden conseguir y, con base en ello, debe seleccionarse una muestra de reactivos para que con ello se pueda cumplir con la meta establecida. Esta selección de ítems también nos va a resultar útil para conocer las propiedades y limitaciones que asumimos cuando aplicamos un determinado cuestionario.

Supongamos que si en la fase de análisis de ítems se tuviera como objetivo elaborar un test con elevada consistencia interna. Entonces, para lograr esto tenemos que quedarnos con los ítems que manifiestan un mayor índice de homogeneidad. Sin embargo, no es tan sencillo como aparenta, puesto que se debe tener en cuenta que el coeficiente alfa aumenta cuando incrementamos la longitud del test y que resultaría fácil obtener valores elevados cuando se incluyen ítems redundantes, lo que, evidentemente,

no resulta deseable. De tal suerte que el error de selección de reactivos se puede limitar con base a la elección precisa de cada constructo en función a la sintaxis y semántica que compone a cada uno de éstos (Magnusson, 1977).

Por otro lado, Martínez et al, (2015), sostiene que el coeficiente de fiabilidad test-retest (Temporal; rxx), su cuantía depende en parte de la variabilidad de la muestra donde se obtiene y también de la longitud (número de ítems) del test. Debemos conocer que un mismo test tiene diferentes rxx en diferentes grupos normativos (muestras de personas donde se obtiene el coeficiente). Más concretamente, un mismo test suele obtener un rxx mayor en un grupo heterogéneo que en otro menos heterogéneo (de menor varianza). Por tal razón se espera que en poblaciones-muestra más grandes las puntuaciones obtenidas sean más elevadas, mientras que en situaciones en las que se evalúen muestras más pequeñas se observarán puntuaciones más bajas con respecto al promedio. Por tal motivo, se establece que en tanto los ítems estén bien formulados y resulten discriminativos, entonces un test incrementará su rxx a medida que se aumente su longitud (número de ítems).

Por otra parte, Aragón (2015), Mendenhall *et al.* (2017) así como Prieto y Delgado (2010), mencionan que otra fuente de error consiste en la aplicación presencial de una prueba. Cada instrumento cuenta con su manual de aplicación, mismo que establece las condiciones bajo los cuales se garantiza la objetividad y cientificidad de las puntuaciones obtenidas a partir de la implementación del test. En éste se pueden encontrar las instrucciones de manera detallada sobre la manera correcta en la que se debe presentar el test (condiciones ambientales, del evaluador y el evaluado), las instrucciones verbales, el papel del evaluador y el evaluado, el control de los resultados e, incluso, las instrucciones sobre cómo realizar un análisis de datos efectivo sobre las puntuaciones recabadas. De manera general se explica aquí las condiciones bajo las cuales el instrumento está estandarizado. Infortunadamente, en la mayoría de los casos no es posible controlar todas las variables interventoras que afectan a nuestra evaluación psicométrica; a éstos factores se les considera parte del error aleatorio de medida.

También el control de las puntuaciones, así como sus calificaciones o valoraciones conforman una parte significativa de los errores de medición. Si bien es cierto que el manual del instrumento cuenta con plantillas, claves y/o con instrucciones detalladas de

calificación de los ítems, en ocasiones las pruebas cuentan con reactivos que tienen que ser valorados por el evaluador; tal es el caso de las preguntas abiertas donde el investigador tiene que decidir si se le otorga una puntuación (y el grado de dicha puntuación) o no. Para combatir este error de medida, los creadores de los test psicométricos postulan ejemplos en los manuales sobre las posibles respuestas y la forma en que deben ser valorados (Abad et al., 2006 y Moreno, 2006).

Todas las características de las muestras con las que se obtienen los estimadores de los coeficientes de fiabilidad deben ser descritas en el manual del test o en el apartado del método en un artículo o trabajo de investigación. Ello a causa de que las prácticas modernas de publicaciones de las investigaciones psicológicas (Wilkinson y APA Task Force, 1999, citados en Martínez, et al., 2014), insisten en que deben exponerse los coeficientes de fiabilidad de las puntuaciones de los instrumentos utilizados. Dada la sensibilidad de los coeficientes a las características de las muestras, sería conveniente que se calculase la fiabilidad de las puntuaciones de cada estudio en concreto.

Con respecto a ello, Martínez, et al., (2014) señala que no existe un valor específico y adecuado para la fiabilidad de las puntuaciones, sin embargo, en lo que va de la práctica, se sugiere seguir las recomendaciones de Nunnally (1978), quien postula una serie de valores mínimos de referencia en función del uso de la prueba. Suelen considerarse aceptables en los trabajos de investigación en los que se usan puntuaciones de test valores ≥ 0.70, pero los valores deben ser mucho más grandes cuando las puntuaciones se utilicen para tomar decisiones sobre sujetos concretos, siendo en este caso el mínimo recomendable de 0.90 y lo deseable de 0.95. También establece no descuidar los coeficientes de validez, en el cual se aplica la misma regla del coeficiente de fiabilidad; es decir, la precisión y consistencia en una medida son siempre deseables. No obstante, esta necesidad aumenta a medida que las consecuencias de las decisiones crecen en importancia (AERA, et al., 1999 citado en Martínez, et al., 2014).

Finalmente, cabe mencionar que Spearman (citado en Martínez, et al., 2014), demostró los efectos de los errores de medida en el análisis de datos de la siguiente manera:

1. En los estudios de comparaciones de grupos: reduce el tamaño del efecto, reduce la potencia estadística de los contrastes y aumenta la varianza de los tamaños del efecto.

2. En los estudios basados en correlaciones, además del efecto de atenuación o reducción del tamaño correlativo, aumenta la variabilidad de la distribución muestral, reduciendo la efectividad; principalmente en coeficientes de regresión y análisis factoriales.

2.4. Teoría de la Generalizabilidad

Este enfoque clásico ha generado diversas variantes sobre todo en función del tratamiento dado al error de medida. Ha habido numerosos intentos de estimar los distintos componentes del error, tratando de descomponerlo en sus partes. De todos estos intentos el más conocido y sistemático es la Teoría de la Generalizabilidad (TG) propuesta por Cronbach y sus colaboradores (Cronbach, Gleser, Nanda y Rajaratnam, 1972). Se trata de un modelo de uso complejo, que utiliza el análisis de varianza para la mayoría de sus cálculos y estimaciones. Esta suele considerarse una extensión de la Teoría Clásica de los Test (TCT), con el cual es posible examinar cómo es que los diferentes aspectos de las mediciones pueden afectar el grado de confianza que podemos tener en las inferencias basadas en las puntuaciones (Brennan, 2001 y Cronbach, et al., 1972, citados en Martínez, et al, 2014 y Zúñiga-Brenes, 2007).

Esto es posible a través del modelo de análisis de varianza (ANOVA), no obstante, de acuerdo con Zúñiga-Brenes (2007), Muñiz, (2010) y Martínez et al., (2014), la TG va más allá que la sola aplicación del modelo. Ello a causa de que esta teoría intenta dar solución a las siguientes limitantes de la TCT:

1. La concepción unitaria e indiferenciada del error de medida. En donde se engloban todos los posibles errores por existir en una evaluación psicológica. No obstante,

esta concepción no resulta ajusta a la práctica de los test y, por tanto, no resulta útil en el diseño de los mismos.

2. La rigidez del concepto de paralelismo de las medidas (tau-equivalentes, esencialmente equivalentes y congeneritas). En donde resulta sumamente complicado mantener estas formas de equivalencia de medidas; un claro ejemplo de ello es el cómo se entiende el concepto de calificadores o jueces paralelos.

3. Polisemia del concepto de fiabilidad. La TCT descompone la varianza observada en verdadera y de error. Y es justamente ésta única concepción de error de medida la que conlleva a crear estimaciones de fiabilidad inscritas en este error de medición indiferenciado que, en función del procedimiento de estimación utilizado, reflejará fuentes de error distintas y tendrás, por tanto, un significado distinto.

Díaz, Botanero y Cobo (2003) y Prieto y Delgado (2010), sostienen que son estas limitantes de la TCT lo que constituye el núcleo de la TG, ya que ésta pretende reducir el margen de error retomando a todos los posibles factores que influyen en las puntuaciones obtenidas de la evaluación de uno o varios participantes; incluyendo, así mismo a los sujetos *per se*, instrumentos y/o condiciones bajo los cuales se ejecutó la evaluación como variables interventoras.

En la teoría G se analiza la variabilidad de los puntajes observados según fuentes separadas de variabilidad. Por ejemplo, en un diseño (p x i) la variabilidad se divide en tres fuentes: personas, ítems y el residuo. Así, la teoría G define los componentes de varianza ($\sigma_2^{\wedge}$) para cada fuente de variabilidad de los puntajes observados. En este caso, estos se denominan el componente de varianza de las personas (σˆ 2 p), ítems (σˆ 2 i) y el residuo (σˆ 2 pi, e) (Zúñiga-Brenes, 2007; Kim y Wilson, 2009).

Una segunda consecuencia de la TG es que amplía con ventaja el cálculo clásico de la fiabilidad. Este modelo es especialmente útil para evaluar la fiabilidad de las calificaciones otorgadas por evaluadores a los productos obtenidos en pruebas o exámenes abiertos (los examinados no están constreñidos por un formato cerrado, tal

como los de las pruebas de elección múltiple, para emitir sus respuestas). Y con base en ello, es que la TG permite calcular al menos dos coeficientes, uno de los cuales (generalizabilidad a otros ítems) coincide con el coeficiente (Díaz et al, 2003).

Es por ello que la TG concibe a la fiabilidad como la exactitud al generalizar un puntaje obtenido por una persona en una prueba u otra medida al puntaje promedio que la persona habría recibido bajo todas las posibles condiciones de medición (Shavelson y Webb, 1991 citado en Zúñiga-Brenes, 2007; Kim y Wilson, 2009). En otras palabras, la fiabilidad es ese coeficiente que está en relación con las diferencias que existen entre las personas, las ocasiones en que se realice la prueba, los observadores o calificadores, los ítems que se utilicen y otras condiciones presentes en el estudio. Así, un solo puntaje obtenido en una ocasión en particular, en una prueba con un solo observador no es totalmente fidedigno.

Zúñiga-Brenes (2007) y Kim, y Wilson (2009) postulan que el desempeño de la persona en cualquier muestra de ítems se podrá generalizar a todos los reactivos siempre y cuando la dificultad de los ítems no varíe. Y es la variabilidad de los mismos lo que representa una fuente potencial de inconsistencia en la generalización; siendo ésta tan sólo una faceta del proceso de la TG. Y en el caso de psicología se requieren regularmente más de dos facetas, por ejemplo, de ítems, observadores, calificadores, contextos, ocasiones (momentos), etc.

Ahora bien, para la creación de cada faceta se requiere tomar en cuenta los siguientes factores propuestos por Shavelson y Webb (1991, citado en Zúñiga-Brenes, 2007)

1. La primera fuente de variabilidad se encuentra en las diferencias sistemáticas entre las personas en el rasgo o constructo que se desea medir; esto es, la variabilidad entre los objetos de medida (normalmente las personas), la cual se refleja en las diferencias de conocimiento, habilidades u otros atributos entre los examinados(as).

2. La segunda fuente de variabilidad es la diferencia en la dificultad de los ítems de la prueba. Algunos reactivos se consideran fáciles, intermedios o difíciles, según su nivel de dificultad, medido empíricamente, por ejemplo, en términos de la proporción de respuestas correctas para un grupo de examinados(as).

3. La tercera fuente de variabilidad se refleja en el nivel educativo y experiencias previas que las personas hayan tenido. Por ejemplo, un ítem de una prueba de ciencias que se refiera a hámster, sería posiblemente más fácil para una persona que los ha tenido o tiene como mascota. Esto implica una interacción entre las personas y los ítems. Este emparejamiento entre las experiencias de una persona y un reactivo en particular, aumenta la variabilidad entre personas e incrementa la dificultad para generalizar, en términos del atributo específico que se desea medir.

4. La cuarta fuente de variabilidad se supone que es debida a otros factores sistemáticos no identificados o no conocidos.

De manera general, cada faceta representa cada una de las características de la situación de medida. Éstas son equivalentes a los factores que usamos en el modelo ANOVA, y sus efectos son considerados efectos principales, mientras que las combinaciones entre facetas pueden analizarse como interacciones y las diversas manifestaciones de estas combinaciones se denominan condiciones. Mismos que a su vez se consideran equivalentes a los niveles de la TG.

Finalmente, con base en las propuestas de innovación que la TG ofrece a la TCT, Martínez et al., (2014) describe las siguientes soluciones a los errores de medición producidas a partir de la TG:

1. La TG utiliza el concepto estadístico de muestreo de fuentes de variación múltiple (en lugar de la concepción de error de medida), ello le permite tratar cada una de las características de la situación de medida como una faceta de un plan de medición, concebido como un diseño experimental. Y, puesto que cada faceta

cuenta con su nivel de variabilidad entonces permitirá considerar su variación por medio del Modelo Lineal General.

2. Sustituye el concepto de “medidas paralelas” por el de “medidas aleatoriamente paralelas”, el cual considera que los distintos componentes de evaluación puedan considerarse como una muestra aleatoria de un universo más amplio, definido por las condiciones de muestreo de cada situación de evaluación.

3. Se amplía el concepto de fiabilidad y éste se convierte en un problema global que conlleva a la generalización o inferencia estadística a un universo a partir de las puntuaciones observadas

    a. El error se concibe como una fluctuación muestral correspondiente a la extracción aleatoria de algunas condiciones.

    b. Para determinar en qué medida un dato observado se aproxima a las puntuaciones del universo se analiza el grado de generalizabilidad o invarianza dentro de la población; de esta manera se sustituye el concepto de fiabilidad por el de generalizabilidad del resultado.

2.5. Limitantes del uso de pruebas psicométricas

Del enfoque de la teoría clásica bien podría decirse que goza de muy buena salud, hay pocas dudas de su utilidad y eficacia. La pregunta obligada aquí es por qué hacen falta otras teorías de los test, o, en otras palabras, ¿qué problemas de medición no quedaban bien resueltos dentro del marco clásico? ¿Cuándo podemos afirmar que un test es justo?

La mayoría de los problemas en torno a los test provienen de su uso inadecuado, más que del test en sí mismo, de su construcción o de sus propiedades técnicas. Muñiz y Hambleton (1996, citado en Gómez-Benito et al, 2010), exponen que los aspectos como el contexto sociocultural, el proceso de construcción y/o adaptación, las condiciones de aplicación, la interpretación de las puntuaciones y el grado de formación del profesional pueden ocasionar que el test sea injusto en su aplicación. Por otro lado, Gómez-Benito et

al. (2010), asumen que las dos primeras cuestiones están solventadas y que, por lo tanto, el interés se traslada a las propiedades técnicas o psicométricas del test. De esta forma, ellos establecen que estas propiedades se dividen en tres: Sesgo, Funcionamiento Diferencial del Ítem e Impacto.

El primero hace hincapié en la existencia de un sesgo cultural en los instrumentos de medida psicológicos puede representar una seria amenaza contra la validez de dichos instrumentos en los que algunos de sus ítems están beneficiando a ciertos grupos de la población en detrimento de otros de igual nivel en el rasgo que interesa medir. El sesgo se refiere a la injusticia derivada de uno o varios ítems del test al comparar distintos grupos que se produce como consecuencia de la existencia de alguna característica del ítem o del contexto de aplicación del test que es irrelevante para el atributo medido por el ítem (Martínez et al., 2014).

Por otra parte, el *Differential Item Functioning* (DIF) adoptado por el Holland y Thayer (1988, citado en Gómez-Benito et al., 2010) afirma que un determinado ítem presenta DIF si a nivel psicométrico se comporta diferencialmente para diversos grupos, es decir, una diferencia del funcionamiento del ítem (o test) entre grupos comparables, entendiendo por comparables aquellos grupos que han sido igualados respecto al constructo o rasgo medido por el test, por ejemplo, el caso de grupos focales y grupos referenciales (Potenza y Dorans, 1995, citado en Gómez-Benito et al., 2010).

Sin embargo, el hecho de que un instrumento de medida obtenga resultados sistemáticamente inferiores en un grupo en comparación a otro no necesariamente implica la presencia de DIF, sino que pueden existir diferencias reales entre los grupos en el rasgo medido por el test en cuestión. En este caso se habla de impacto o diferencias válidas (Van de Vijver y Leung, 1997, citado en Gómez-Benito et al., 2010).

Ahora bien, Martínez, et al., (2014) y Muñiz, (2010) establecen que dichas propiedades dentro del marco de la TCT conllevan a una alteración de las mediciones, lo cual no resulta invariante respecto al instrumento utilizado. Por ejemplo, si un psicólogo evalúa la inteligencia de tres personas distintas con un test diferente para cada persona, los resultados no son comparables, no podemos decir en sentido estricto qué persona es más inteligente. Esto es así porque los resultados de los tres test no están en la misma

escala, cada test tiene la suya propia. Para hacerlo se transforman las puntuaciones directas de los test en otras baremadas, por ejemplo, en percentiles, con lo que se considera que se pueden ya comparar, y de hecho así se hace. Este proceder clásico para solventar el problema de la invarianza no es que sea incorrecto, pero, amén de poco elegante científicamente, descansa sobre un pilar muy frágil, a saber, se asume que los grupos normativos en los que se elaboraron los baremos de los distintos test son equiparables, lo cual es difícil de garantizar en la práctica.

Otra gran cuestión no resuelta del todo dentro del marco clásico era la ausencia de invarianza de las propiedades de los test respecto de las personas utilizadas para estimarlas. En otras palabras, propiedades psicométricas importantes de los test, tales como la dificultad de los ítems o la fiabilidad del test, estaban en función del tipo de personas utilizadas para calcularlas, lo cual resulta inadmisible desde el punto de vista de una medición rigurosa. Por ejemplo, la dificultad de los ítems, o los coeficientes de fiabilidad dependen en gran medida del tipo de muestra utilizada para calcularlos. (Mendenhall et al 2017; Usabiaga, Castellano, Blanco-Villaseñor y Casamichana, 2013; Zúñiga-Brenes, 2007).

Aparte de estas dos grandes cuestiones, había otras menores de carácter más técnico a las que la teoría clásica no daba una buena solución. Por ejemplo, cuando se ofrece un coeficiente de fiabilidad de un test en el marco clásico, como el coeficiente alfa de Cronbach, se está presuponiendo que ese instrumento mide con una fiabilidad determinada a todas las personas evaluadas con el test, cuando tenemos evidencia empírica más que suficiente de que los test no miden con la misma precisión a todas las personas, dependiendo la precisión en gran medida del nivel de la persona en la variable medida (Muñiz, 2010).

Y, finalmente, Mendenhall et al (2017), Muñiz (2010), Gómez-Benito et al. (2010) y González (2007), sostienen que para poder resolver los problemas citados anteriormente que no encontraban una buena solución dentro del marco clásico, nace una nueva teoría que busca resolverlos con base a asunciones más fuertes y restrictivas que las hechas por la Teoría Clásica. Dicha teoría se le conoce como “Teoría de Respuesta al Ítem” (TRI).

Y, de acuerdo con Martínez (2014), el supuesto clave en los modelos de TRI es que existe una relación funcional entre los valores de la variable que miden los ítems y la probabilidad de acertar a estos, denominando a dicha función Curva Característica del Ítem (CCI). El segundo supuesto describe que la mayoría de los modelos de TRI, y desde luego los más populares, asumen que los ítems constituyen una sola dimensión, por tanto, antes de utilizar estos modelos hay que asegurarse de que los datos cumplen esa condición. Por último, el tercer supuesto de los modelos de la TRI es la denominada Independencia Local, que significa que para utilizar estos modelos los ítems han de ser independientes unos de otros, es decir, la respuesta a uno de ellos no puede estar condicionada a la respuesta dada a otros ítems. En realidad, si se cumple la unidimensionalidad también se cumple la Independencia Local, por lo que a veces ambos supuestos se tratan conjuntamente.

3.ESTRATEGIAS DE APLICACIÓN DE LAS PRUEBAS PSICOMÉTRICAS

El desarrollo y aplicación de las Tecnologías de la Información y Comunicación (TICs), han impactado en la forma de vida de la sociedad en general. La implementación de las mismas va desde las actividades de entretenimiento hasta su uso en aplicaciones científicas. Indudablemente, la psicología no es ajena a esta tendencia; aquí la investigación científica requiere de la captura y procesamiento de grandes cantidades de información. Razón por la cual los recursos humanos y el tiempo necesario para realizar estas actividades dependen del tamaño de la muestra y del tipo de instrumento utilizado; es por ello que las herramientas digitales comprenden una parte fundamental para el desarrollo y consolidación de la psicología como ciencia.

3.1. Medios sistematizados de aplicación de los test psicológicos

A partir de los pilares que impulsaron el desarrollo de las pruebas psicométricas a finales del siglo XIX y a principios de siglo XX fue que se mantuvo una constante preocupación no sólo por la construcción y estandarización de las pruebas, sino el proceso de aplicación a una población determinada que permitiese economizar y agilizar el proceso de evaluación y, por lo tanto, de toma de decisiones. Es por ello que, a partir del año de 1960, en el origen de la segunda generación de computadoras, comenzó a utilizarse las técnicas de automatización en salud mental para evaluación a pacientes clínicos; ello dio origen a la creación de pruebas psicológicas sistematizadas o computarizadas; siendo las pruebas de personalidad las primeras evaluaciones en volverse informatizados (Molinar et al., 2012).

Los Test Adaptativos Informatizados (TAI) estrictamente hablando, son aquellos que utilizan la computadora como medio de presentación de sus ítems, el registro de las respuestas y de análisis e interpretación de los rendimientos (Olea, Ponsada y Prieto, 1999 citado en Aguilar-Espinoza, 2013).

De esta manera surgieron diversos esfuerzos por implementar estas pruebas psicométricas a través de las nuevas tecnologías, un ejemplo de ello fue el trabajo de Priále en (1983, citado en Aguilar-Espinoza, 2013), quien realizó una presentación sobre las pruebas psicológicas computarizadas en el Perú y, de manera inmediata en el año de 1984, se realizaban ya más de 300.000 interpretaciones computarizadas de test psicológicos anualmente.

Posteriormente Escurra, Delgado y Aparcan (1986, citados en Aguilar-Espinoza, 2013) ilustraron el uso de la computadora bajo el enfoque de la Teoría de Respuesta al Ítem (TRI) al presentar su estudio sobre el modelo Rasch como un caso especial del uso de estructuras latentes. Obteniendo de esta manera que las características del modelo de Rasch referidas a la dicotomía de los ítems, la monotonicidad, la unidimencionalidad, la independencia local y la suficiencia estadística de la simple suma de los resultados, ilustra la necesidad del uso de computadoras y los programas respectivos para su pertinente aplicación.

Fue así que los primeros TAIs elaborados en el mundo se usaron en la milicia y fueron desarrollados por ASVAB en los Estados Unidos y MicroPAT en Europa. En la actualidad, el uso de TAIs se ha extendido a varios test estandarizados como el Test of English as a Foreing Language (TOEFL) del ETS y a distintos campos de evaluación como aptitudes, personalidad y conocimientos (Aguilar-Espinoza, 2013).

Aguilar-Espinoza (2013) y Morales-Ramírez y cols., (2012), realizaron un recorrido histórico sobre los diferentes avances logrados a raíz de la aplicación de test psicométricos a través de medios digitales. Ellos describen este proceso de la siguiente manera:

- Año 2000: Schuhfried presentó un excelente catálogo sobre pruebas psicológicas computarizadas, destacando el sistema de pruebas Viena, el cual incluía más de cien pruebas psicológicas digitales; un año después Collins y Sayer desde una vertiente de la investigación ofrecen información sobre la aplicación de programas informáticos que apoyan al sector laboral en la aplicación de estos test:

- Año 2003: Weber y colaboradores analizaron la atención y la memoria en pacientes psiquiátricos mediante la aplicación de pruebas psicológicas por computadora.

- Año 2005: Stoddard *et al.,* evaluaron un tratamiento para dejar de fumar mediante el internet; en este mismo año 2005 Niemz, Griffiths y Banyard valoraron el uso psicopatológico de Internet a través de la Web.

- Año 2006: Christensen, Griffiths, Mackinnon y Brittliffe evaluaron un tratamiento cognitivo-conductual diseñado para disminuir la depresión.

- Año 2007: González-Santos, Mercadillo, Graff y Barrios aplicaron la versión computarizada del listado de síntomas 90 (SCL 90) y del inventario de temperamento y carácter (ITC) para realizar un estudio psicopatológico; dando pie con ello a la introducción de la medicina, psicología y psiquiatría a la evaluación digital.

- Año 2012: Un avance importante en este sentido de las soluciones tecnológicas aplicadas fue presentada por Molinar-Solís y cols., (2012), quienes reportaron el desarrollo de una aplicación que, mediante el procesamiento digital de la hoja de respuestas, posibilita la evaluación automática o sistemática de la prueba del Inventario Multifásico de la Personalidad de Minnesota, evitando de esta manera la calificación manual del instrumento y, logrando de esta forma, la disminución de costos y aumento de productividad.

Finalmente cabe señalar que autores como Tirado, Backhoff y Larrazolo (2016) señalan que a través de la última década no sólo se ha producido una inflexión en el diseño de programas computarizados, sino que además, la manera en que la información es introducida o retirada en el ordenador sufre una importante modificación; hasta hace unos años el sistema más utilizado era el denominado *batch processing* (la información es recogida en impresos, para posteriormente transmitir los datos al instrumento digital), mientras que hoy en día se prefiere el sistema online, en donde los datos son suministrados (y retirados) directamente del instrumento digital. Por todo ello, en la última década se han abandonado los grandes ordenadores en favor de los más disponibles, baratos y

manejables computadores personales, los cuales, permiten el acceso a "softwares" comercializados que obvian la elaboración de programas propios para el sistema informatizado.

3.2. Tecnologías de la información y la comunicación (TIC)

Antes de comenzar a abordar de lleno los procedimientos que conformaron las bases de la aplicación digital de las pruebas psicológicas es necesario hacer una distinción entre éstas, es decir, discernir entre cuales son las pruebas psicométricas de primera y segunda generación. Para ello, Aguilar-Espinoza (2013) define a los Test Informatizados de primera generación como aquellos test que existieron por vez primera en formato de lápiz y papel; y también se incluyen en esta categoría a aquellos test que, aun siendo desarrollados para su administración computarizada no incluyen algoritmos complejos para la selección de los ítems, la presentación de los mismos o el tratamiento de las puntuaciones. Ahora bien, dentro de los Test de segunda generación o test Adaptativos Informatizados y, a diferencia de los Test de primera generación, existe una elección del ítem siguiente de acuerdo al proceso, es decir, en lugar de utilizar un algoritmo matemático complejo se pide o se somete al sujeto evaluado a que elija (o no, de acuerdo a su ejecución y prueba) el nivel de dificultad del ítem; con la finalidad de que el control que tiene el sujeto sobre la situación de evaluación minimice el efecto negativo que la ansiedad tiene sobre una prueba de ejecución. Razón por la cual los Test Adaptativo Informatizado (TAI) se definen como un instrumento conformado por un banco de ítems, calibrado por los principios de la Teoría de Respuesta al Ítem (TRI), que implica un procedimiento para la estimación del nivel de habilidad del examinado y otro para seleccionar el ítem más adecuado de acuerdo a dicho nivel y cuya elaboración, aplicación, calificación y actualización se realiza por medio de un soporte informático (Hambleton, Zaal y Pieters, 1991, citado en Sierra–Matamoros et al, 2007).

Sierra–Matamoros *et al.* (2007), establecen que el propósito principal de la aplicación de los TAIs es reducir el tiempo de administración de un test; el punto de partida es que cada sujeto sea evaluado simplemente con los ítems que proporcionan máxima información para su nivel de aptitud a partir de un soporte digital que permita su cuantificación objetiva a través de proceso estadísticos. Así, cada sujeto puede ser

evaluado con un conjunto diferente de ítems. La calibración de los ítems bajo uno de los modelos de la TRI tiene como propósito la estimación de los parámetros de los ítems, teniendo en cuenta la invarianza de la medida respecto del instrumento y del grupo de examinados. Dado que la TRI intenta buscar mediciones invariantes respecto del instrumento y de los examinados, su empleo en la construcción y soporte digital de los TAIs es fundamental.

Dicho soporte digital permite llevar a cabo los procedimientos matemáticos y estadísticos requeridos, acceder al banco y presentar el ítem de forma inmediata. El software que se utiliza en los TAIs se caracteriza por constar de una serie de módulos que procesan diferentes pasos de la prueba de forma independiente y se encuentran en una relación jerárquica. Así, se encuentran módulos que permiten la construcción de ítems, la presentación de los mismos, la finalización de la prueba, el cálculo del nivel de habilidad, el almacenamiento de los resultados, la baremación, la actualización del banco de ítems y la presentación de un informe escrito al examinado sobre su desempeño (Olea y Ponsoda, 1996). Entre el software utilizado se encuentra el MicroCAT de la *Assesment System Corporation* (Hambleton et al., 1991; Olea y Ponsoda, 1996; Muñiz, 1997), el DEMOTAC y APT-System, elaborado en España (Olea y Ponsoda, 1996 citado en Sierra–Matamoros et al. 2007).

Las investigaciones sobre la elaboración, el mantenimiento y la renovación de los bancos de ítems se han enfocado, por un lado, a la estimación de parámetros de los ítems, ya sea a partir de los índices psicométricos de la Teoría Clásica de los Test (TCT), la reestimación de los ítems (ítems operativos) y la estimación de nuevos ítems (ítems pretest) y, por otro lado, a la generación automática de ítems, los cuales pueden ser isomorfos, es decir, que posean contenido y propiedades psicométricas similares (Molinar *et al.*, 2012). Así mismo, se han propuesto TAIs de razonamiento cuantitativo, por Bejar, Lawless, Morley, Wagner, Bennett y Revuelta y TAIs basados en un modelo de respuesta multinivel, por Glass y van der Linden (2003, citados en Molinar *et al.*, 2012).

También se han investigado las propiedades de los TAIs por medio de procedimientos de simulación (Chang y Ying, 2004; Olea y Ponsoda, 1996; Abad, Olea, Ponsoda, Ximénez y Mazuela, 2004). Así, por ejemplo, en el estudio de Abad et al. (2006), se encontró que si un banco de ítems calibrado trata las omisiones como

respuestas fraccionalmente correctas produce valores estimados de habilidad más parecidos a la habilidad verdadera del sujeto, a diferencia de cuando se tratan las omisiones como errores, que conducen a la sobreestimación de los niveles de habilidad.

Olea y Ponsoda (1996, citados en Abad et al., 2006), sugieren que la investigación sobre TAIs se llevará a cabo en torno a tres temas principales: Condiciones de aplicación, implementación de nuevos modelos de TRI y extensión de los contenidos a evaluar. Estas temáticas se refieren a asuntos como el uso de bancos cortos o la posibilidad de revisar o diferir las respuestas, superar la restricción de unidimensionalidad y evaluar procesos psicológicos básicos (memoria, percepción, atención) respectivamente.

3.3. Medios digitales de aplicación.

Partiendo de la idea de que el objetivo principal de la administración de un test digital es la reducción de costes y la cuantificación objetiva del mismo, fue como nacieron diversos programas que se encargan de satisfacer estas necesidades de todo evaluador. Por ejemplo, González y Hernández (2013) expusieron en su trabajo la eficiencia del uso del Lenguage Visual Basic 10, al cual describen como un programa adaptable a los gráficos de la batería que el test requiere y su uso comercial es fácil de adquirir e implementar.

Los autores reportaron que el haber alcanzado una interfaz gráfica (descrita por los usuarios) del 86.2% en la evaluación postest fue posible gracias al uso de este software. Ahora bien, referente a la recaudación de los datos, éste mostró en un 98.8% de almacenamiento y la obtención de datos total fue de 100%, lo cual permitió conservar la validez y confiabilidad de los mismos. Esto indica que la sistematización de la batería del test fue exitosa y ágil. Concluyendo de esta manera que, de acuerdo con el diagnóstico obtenido a partir de la batería digitalizada se pudo entonces elaborar estrategias de aprendizaje y evaluación adecuadas a cada uno de los perfiles a los que se pretendía implementar dicha labor.

Por su parte, Morales-Ramírez et al., (2012), describen el uso de un sistema específico que se desarrolló, implementó y administró a través del lenguaje de

programación PHP y con ayuda de las herramientas MySQL; el cuales un software open source que proporciona un servidor de base de datos *Structured Query Language* muy rápido de phpMyAdmin; y Adobe Dreamweaver, misma que permitió diseñar, desarrollar y realizar el mantenimiento de aplicaciones y sitios web de gran calidad basados en estándares y Microsoft IIS (*Internet Information Services),* ofreciendo de esta manera un servidor web para el desarrollo, implementación, hospedaje y administración de sitios web.

Todos estos instrumentos tecnológicos se llevaron a cabo en el estudio de Morales-Ramírez et al (2012), con el único propósito de aplicar y gestionar la prueba psicométrica conocida como el Cuestionario Honey-Alonso de Estilos de Aprendizaje (C. H. A. E. A.) (Alonso, Gallego y Honey, 1997, citado en Morales-Ramírez et al., 2012). El cual tiene como objetivo evaluar el estilo preferido de aprendizaje y consta de 80 ítems agrupados en cuatro subescalas (activo, reflexivo, teórico y pragmático), con dos opciones de respuesta Más (+) y Menos (-).

En dicho estudio se obtuvo como resultado que el tiempo que los participantes invirtieron para responder la versión computarizada de la prueba C. H. A. E. A. fue de entre 15 y 20 minutos, tiempo similar al requerido cuando éste es aplicado en su versión lápiz-papel, sin embargo, bastaron 20 minutos para aplicarlo a más de 255 alumnos de manera simultánea mientras que, por otro lado, si la aplicación hubiese sido a lápiz y papel habría requerido de 6 días en jornadas de 6 h para evaluar a toda la muestra poblacional. Lo cual denotó su alta eficacia en cuanto al tiempo que los participantes invierten al responder una prueba en el sistema digital y el tiempo que les toma hacerlo en la versión lápiz-papel.

En este mismo sentido, Butcher y cols., 2004; González y cols., 2007; Hays y McCallum, 2005; Parkin, 2000; Sahakian y Owen, 1992, citados en Molinar, Escoto, García y Bautista (2012), sostienen que las pruebas que se basan en el uso exclusivo de la técnica de lápiz-papel tienen varias diferencias importantes con sus contrapartes computarizadas, según lo establecido en diferentes trabajos. Se pueden resumir tales diferencias en la siguiente tabla:

Tabla 1

Aplicación de Test Psicométricos digitales y a lápiz y papel.

Técnica	Tiempo para responder la prueba	Calificación obtenida	Accesibilidad	Tiempo en calificar	% de error
Lápiz-papel	Similar	Idéntica	Alta	Alto	Alto
Computarizada	Similar	Idéntica	Baja	muy Bajo	Muy Bajo

Tabla 1: Diferencias generales de la aplicación de test psicométricos digitales y a lápiz y papel (Molinar et al, 2012).

Ello fue la razón por la cual los autores propusieron la aplicación de un algoritmo para la evaluación de pruebas psicológicas suministradas con la técnica de lápiz-papel por medio del reconocimiento de patrones empleando el paquete MATLAB® (The Mathworks Inc., 2001). El algoritmo, a diferencia de aquellos programas que utilizan lectores ópticos especiales, solamente requiere de un archivo que contenga la imagen digitalizada de la hoja de evaluación de la prueba en formato bmp (bit-map). En este caso, a manera de prueba, se diseñó una hoja de respuestas con los reactivos correspondientes del MMPI-2 para aplicar el algoritmo. Volviendo así a cualquier computadora personal convencional provista de un software con el algoritmo y de un escáner una herramienta eficaz para realizar la evaluación automática de cualquier prueba psicológica que utilice ítems de opción múltiple.

Se obtuvo como resultados que el tiempo de procesamiento para la obtención de la gráfica del perfil a partir de la resolución de los 567 reactivos de MMPI-2 fue de tan sólo 40 segundos y que el error de clasificación fue menor al 0.5%, y ello se debió a aquellos cuadros que no fueron rellenados adecuadamente, por lo que no fueron

reconocidos por el elemento estructurante del escáner. Dicho estudio nos permite observar a detalle que MATLAB® fue una herramienta importante para lograr dichos resultados, razón por la cual los autores sugieren que éste debe ser considerado y desarrollado en el ámbito de la psicometría. Además, gracias a éste se pueden generar los resultados en formatos electrónicos, lo que haría posible generar un acervo extenso de datos poblacionales y, de esta manera, la creación de importantes bases de datos.

Es de esta manera Muñiz (1997, citado en Eiroá, Fernández y Pérez, 2008), sintetiza la teoría del uso del PC para presentar ítems complejos tales como mediciones de tiempos de respuesta, sonido, simulación y conducta interactiva propuesta por. Ahora bien, hoy en día se conoce que otra ventaja práctica de la aplicación digital de las pruebas es la agilidad en la calificación, gracias a la utilización de un software especializado; dicha calificación no tiene en cuenta el número de aciertos y errores sino los valores de los distintos parámetros de los ítems de acuerdo con el modelo de TRI y las estrategias diversas estrategias de evaluación. Sin embargo, la aplicación de los TAIs puede resultar más costosa que la aplicación de los test clásicos de lápiz y de papel debido al requerimiento de equipos y de software especializado. Además, los examinados pueden experimentar cierto nivel de ansiedad al responder el Test Informatizado, bien sea porque no sienten control sobre el proceso o porque no pueden realizar ciertas correcciones en respuesta de las que no están seguros; ante lo cual se pueden emplear test autoadaptativos informatizados (TAIs), en los que el examinado va determinando su nivel de dificultad para reducir su nivel de ansiedad.

En otras investigaciones citadas en la revisión teórica de Eiroá et al. (2008) que tenían como objetivo la observación y descripción del comportamiento de los resultados de una evaluación digital en contraposición de una evaluación a lápiz y papel, diversos autores como Hallfors, Khatapoush, Kadushin, Watson, Saxe (2000), pretendían saber si la administración de un cuestionario sobre consumo de drogas a través de computadora denotaría un nivel mayor de consumo a diferencia del uso de métodos tradicionales, pero el nivel de consumo declarado no varió. En ese mismo año Vispoel (2000), expuso algunos resultados que apoyaban la comparabilidad de los resultados en relación a la preferencia de la resolución a lápiz-papel o internet del Self-Description Questionnairey, siendo el uso de la versión computarizada la más aceptada por los participantes. Un año

después Vispoel., Boo y Blieiler (2001) describieron una serie de resultados que denotaban pocas diferencias en la estructura psicométrica de la escala de autoestima de Rosenberg según su aplicación vía web o papel. En ese sentido, Epstein, Klinkenberg, Wiley, y McKinley (2001), utilizaron dos muestras equivalentes para comparar la implementación en papel e Internet de un cuestionario sobre atractivo. En este caso no se encontraron diferencias en la muestra total, aunque sí se encontraron al examinar los datos por género. Posteriormente, Cronk y West (2002), reportaron que al administrar un cuestionario sobre visión de la moralidad en cuatro condiciones diferentes (Web o papel, en clase o en casa) no se encontraron diferencias significativas en las condiciones experimentales utilizadas. Sin embargo, se hallaron diferencias en la tasa de respuesta según el método. Años después en la investigación realizada por Fortson, Scotti, Del Ben y Chen (2006), en la que se utilizó un diseño factorial 2 x 2 x 2 de medidas repetidas siendo las variables dependientes factores relacionados con trauma, se demostró que la estructura psicométrica apenas variaba.

Ello puso de manifiesto la preocupación de los psicólogos sobre sí es posible no sólo evaluar a las personas a través de test psicológicos, sino también con el uso en conjunto de entrevistas digitales que permitan realizar un diagnóstico más completo y objetivo. Ello fue lo que impulsó el estudio de Carlbring et al. (2002, citado en Campos, Quero, Bretón, Riera, Mira, Tortella y Botella, 2015) en donde se comparó la concordancia entre una entrevista diagnóstica aplicada a través de Internet (CIDI-SF: Composite International Diagnostic Interview Short-Form) (Kessler, Andrews, Mroczek, Ustun y Wittchen, 1998) con la entrevista SCID (First, Spitzer, Gibbon y Williams, 1999) y el test para evaluar el trastorno de pánico (versión larga de la CIDI); mismos que fueron administrados por los autores a una muestra de 53 participantes. Estos completaron la entrevista computarizada a través de una página web dos días antes de recibir la entrevista cara a cara con el clínico. La concordancia obtenida entre ambas entrevistas fue baja (Kappa de Cohen <0,40) para los distintos trastornos incluidos (depresión mayor, trastorno de ansiedad generalizada, fobia específica, fobia social, agorafobia, ataques de pánico y trastorno obsesivo-compulsivo). Sin embargo, con respecto a la prueba psicométrica CIDI se obtuvo un coeficiente de Kappa mejor (κ=0,48) con un porcentaje de acuerdo del 75%. Ello puso de manifiesto que en la evaluación de determinados trastornos psicológicos se obtiene concordancia cuando se compara junto con la

administración de instrumentos a través de internet y aquella entrevista realizada de forma tradicional (Carlbring et al., 2007; Hedman et al., 2010; citados en Campos y cols., 2015).

En lo que va al estudio realizado por Campos y cols. (2015) consistió en llevar a cabo un estudio exploratorio sobre la concordancia entre la evaluación mediante un *screening* a través de Internet y la evaluación tradicional aplicada por el terapeuta cara a cara para la Fobia a Volar a través del software llamado *Screening* SIN MIEDO Airlines (SMA), para la cual se elaboraron preguntas se basadas en la Entrevista Clínica Estructurada por el DSM-IV (First Spitzer, Gibbon y Williams, 1999). Este instrumento incluyó 17 preguntas sobre fobia a volar (3 ítems); claustrofobia (2 ítems); trastorno de pánico (2 ítems); agorafobia (2 ítems) o acrofobia (2 ítems) y síntomas depresivos (6 ítems). Mismo que se comparó con la aplicación tradicional de: I) La entrevista semi-estructurada de la entrevista clínica para los Trastornos del eje I del DSM-IV (SCID-I: Structured Clinical Interview for DSM Disorders) (First et al., 1999); II) Inventario de depresión de Beck segunda edición (BDI-II: Beck Depression Inventory; Beck, Steer y Brown, 1996); III) Cuestionario de opinión sobre el procedimiento de evaluación. Este instrumento fue elaborado por el grupo de investigación Labpsitec específicamente para el presente trabajo. Comprende un total de tres apartados en que los participantes evalúan: 1) la facilidad o dificultad que han encontrado en la evaluación realizada a partir de una escala tipo Likert de 11 puntos (siendo 0 “extremadamente fácil” y 10 “extremadamente difícil”); 2) el grado de acuerdo con una serie de 5 afirmaciones que recogen la opinión sobre el método de evaluación y 3) cómo se han sentido los participantes durante dicho proceso en un conjunto de 5 afirmaciones (Para la segunda y tercera parte de este cuestionario de opinión, los participantes respondieron en una escala tipo Likert de 7 puntos según su grado de acuerdo donde 1 significaba “en absoluto” y 7 “totalmente de acuerdo”) y IV) Cuestionario de preferencias acerca de la evaluación el cual comprende un total de 6 preguntas que rastrean las preferencias de los participantes acerca de cada uno de los métodos de evaluación, mediante los ítems se valora si los participantes se han sentido cómodos, acogidos, seguros, comprendidos y en qué medida valoraban que su intimidad se respetaba.

Los resultados de dicho estudio demostraron que en lo que va de la facilidad o dificultad informada por los participantes acerca del método de evaluación, se obtuvo una

media de 1,06 (dt=1,59) para la evaluación mediante el programa SMA y una media de 1,33 (dt=1,812) para la evaluación tradicional aplicada por el terapeuta. En la diferencia de medias calculadas para estas puntuaciones, se obtuvo una t=0,898 (p=0,373) que reveló que no existieron diferencias significativas. Ahora bien, en lo que respecta a la opinión recogida sobre la confianza y claridad con el método de evaluación se calculó la diferencia de medias para cada uno de los ítems, los resultados pusieron de manifiesto que no existían diferencias significativas entre la valoración que los participantes hacían de la evaluación aplicada por el terapeuta frente a la evaluación mediante SMA.

Por último, en cuanto a cómo se habían sentido los participantes con ambos métodos de evaluación se obtuvo que la aplicación por el terapeuta contó con puntuaciones medias mayores y significativas respecto a la evaluación mediante el programa. Finalmente, los resultados obtenidos relativos a las preferencias de los participantes los resultados mostraron que el 11,8% elegirían el procedimiento de evaluación SMA frente al 88,2% que elegirían la evaluación aplicada por el terapeuta. Las diferencias entre estos porcentajes fueron significativas, con un valor de χ^2= 24,82, p< 0,001 y una preferencia mayor hacia la evaluación aplicada por el terapeuta. Ello permitió concluir que el programa online proporciona una evaluación válida para la Fobia a Volar. Sin embargo, los resultados obtenidos para la evaluación de otros problemas relacionados como el trastorno de pánico, la agorafobia y la acrofobia fueron menos prometedores. Por otra parte, en lo que va del cuestionario BDI-II (Beck Depression Inventory-II) los resultados mostraron una alta correlación entre las puntuaciones totales obtenidas en formato tradicional y las preguntas incluidas en el *screening* del programa (a pesar que no se comparó todo el instrumento de manera física y digital).

3.4. Internet como medio actual de administración de pruebas psicométricas.

La revolución digital ha dado lugar a una nueva era en la evaluación educativa a gran escala. Junto con los avances de las ciencias cognitivas y la psicometría fue posible lograr avances sustanciales, como la generación automática de reactivos. Fue de esta manera que los adelantos tecnológicos de la informática han permitido desde hace años la

aplicación de pruebas psicológicas a través de internet (Tirado, Backhoff y Larrazolo, 2016)

Se estima en la actualidad que la evaluación a través de Internet puede ser una forma eficaz de recoger información de los usuarios en la investigación psicológica (Vallejo, Jordán, Díaz, Comeche y Ortega, 2007; citado en Campos y cols., 2015). El tipo de test online permite que toda la información (tanto el test como los algoritmos de presentación y los resultados) se almacenen y distribuyan desde un servidor, lo que permite un mayor control sobre los procesos de aplicación y una información inmediata sobre los resultados; establecer enlaces en línea de alta velocidad para transmitir flujos de información en tiempo real y diferido; usar recursos multimedia para hacer presentaciones de textos, imágenes, sonido y video en formatos de alta definición; articular vínculos dinámicos como hipertexto (enlaces internos) e hipervínculo (enlaces web) que permiten relacionar distintas fuentes de información de manera instantánea, evaluar la interacción (respuestas en pantalla); más específico en tanto a los test, permiten obtener una manera controlada y estandarizada de presentar los estímulos, posibilidad de ramificar los ítems (ítem-branching), inmediatez en la entrada de datos, eliminación de respuestas omitidas, eliminación de errores de trascripción, disminución del sesgo del investigador y pueden generar preguntas y procesar las respuestas por algoritmos en cuestión de décimas de segundo (Musch y Reips, 2000, citados en Eiroá et al, 2008; Campos y cols., 2015;Tirado et al., 2016).

Así como la conexión mediante el internet ha representado importantes beneficios logísticos también algunas de estas permiten mejorar las ventajas que ofrecen los ordenadores (Emmelkamp, 2005; Anderson y Titov, 2014; citados en Campos y cols., 2015).

Algunas de estas ventajas son que ofrece *feedback* al terapeuta o evaluador, reduce costos, permite la evaluación y el reclutamiento online, mejora el acceso al tratamiento, llegando a aquellas personas que necesitan ayudan y que no cuenten con las posibilidades de trasladarse. En este sentido, la terapia asistida por ordenador y los tratamientos aplicados a través de Internet han mostrado ser herramientas útiles para diseminar los tratamientos psicológicos. Mostrándose igual de eficaces que los tratamientos “cara a

cara" en una diversidad de trastornos o problemas psicológicos (Anderson, Cuijpers, Carlbring, Riper y Hedman, 2014; citado en Campos y cols., 2015).

Además, permite controlar que el "cliente" (por ejemplo, la empresa o institución que demanda la aplicación) tenga acceso únicamente a la información que resulte pertinente. También existe la posibilidad de evaluar a grandes muestras de participantes, economizar tiempo por parte del psicólogo, evaluar a personas de diferentes lugares, evaluar sin limitaciones de horario, automatizar ítems con la ventaja de que el acceso y manipulación que se haga de los datos sea más fácil e inmediato (Carlbring et al., 2007; Vallejo Jordán, Díaz, Comeche y Ortega, 2007; citado en Campos y cols., 2015). Sin embargo, el uso de internet como medio de transporte de las pruebas psicológicas y de las respuestas de los evaluados requiere tener en cuenta algunas consideraciones en relación a varios riesgos:

- *Calidad:* Cualquiera puede acceder a centenares de test que se ofrecen en todo el mundo y de los que desconocemos sus propiedades psicométricas.
- *Seguridad:* Un importante problema es el de la seguridad del propio test, sobre todo cuando las puntuaciones en los test tienen importantes consecuencias para los evaluados (admisión a un centro educativo, a un puesto de trabajo, acreditación profesional, etc.).
- *Control:* Otro problema importante tiene que ver con las posibilidades de suplantación de identidad, es decir, que sean otras personas las que respondan al test. Una posible solución sería la aplicación controlada por supervisores que aseguren la identidad de los evaluados, que asignen las contraseñas oportunas de acceso y que controlen el cumplimiento de las condiciones de aplicación.

De acuerdo con Morales-Ramírez et al (2012), las versiones computarizadas de pruebas psicológicas, con respecto a la técnica convencional de lápiz y papel, presentan diferencias importantes tales como:

1. Permite una rígida estandarización de instrucciones, tiempos de administración, presentación de ítems, registro de respuestas, corrección y puntuación de las pruebas.

2. Facilita la toma de mediciones simultáneamente en varias variables: acierto/error, latencia de respuesta a cada pregunta, etc., siendo de suma importancia para el desarrollo de test aptitudinales a partir del enfoque del procesamiento de la información (Ronning, Conoley, Glover y Witt, 1987).

3. Admite el almacenamiento, puntuación y análisis estadístico, de los datos sin etapas intermedias de codificación, grabación. Facilitan y agilizan el cálculo de la puntuación de cada subescala (González-Santos, Mercadillo, Graff y Barrios, 2007).

4. Posibilita la utilización de nuevos modelos de pruebas psicométricas, como los test de rasgo latente (Weiss y Davison, 1981; Hambleton y Swaminatham, 1985). Mantienen la flexibilidad en el cambio de los elementos de una prueba (Vispoel, 2000).

5. Permite el diseño y el empleo de test adaptados al sujeto (Weiss y Vale, 1987). Es decir, facilita la ramifican automática los ítems (Eiroá, Fernández y Pérez, 2008; Van de Looij-Jansen; Jan de Wilde, 2008). Es decir, puede iniciar con la presentación de un solo ítem y, de acuerdo con la respuesta dada al mismo, la presentación de un segundo ítem adecuado cuya respuesta llevará a un tercero, y así sucesivamente.

6. Hacen que las personas se sientan más confortables cuando tienen que responder a preguntas confidenciales (Lucas, Mullins, Luna y McInroy, 1977; Greist y col, 1974; Greist y Klein, 1980).

7. Reducen respuestas omitidas (Carine, Vereeckenv y Lea, 2006; Eiroá, Fernández y Pérez, 2008) y Eliminan errores humanos que se cometen en la evaluación manual (Musch y Reips, 2000; González-Santos, Mercadillo, Graff y Barrios, 2007; Cohen y Swerdlik, 2001).

3.4.1. Limitaciones

Las pruebas Psicológicas Sistematizadas presentan al igual que las pruebas Psicológicas de lápiz y papel algunos riesgos; estos riesgos no pueden oscurecer las importantes

ventajas que tienen para el psicólogo las Pruebas Psicológicas Sistematizadas. A continuación Eiroá et al. (2008) puntualiza algunos de los riesgos más importantes que se deben de considerar.

3.4.1.1. Enfatiza que las Pruebas Psicológicas Sistematizadas no son por sí mismas garantes de mejores mediciones.

La validez. Los nuevos formatos de ítems deben ir acompañados de evidencias empíricas de validez. A pesar de que estas pruebas pueden tener varias versiones, sufren el desgaste natural al usarse repetida e intensivamente, como es el caso de las pruebas de ingreso a las universidades, porque se filtran las preguntas, los aspirantes se van aprendiendo las respuestas y en consecuencia pierden su validez. También las respuestas de opción múltiple tienen el inconveniente de ser fáciles de copiar, además de que pueden ser respondidas correctamente al azar. Es por estas razones que los ítems de estas pruebas deben renovarse en forma constante, con el objetivo de mantener vigente la validez de sus resultados. Según Krantz y Dalal (2000 citados en Eiroá et al. 2008) hay dos formas principales de establecer la validez del método de investigación en Internet. En primer lugar, compararlo con otro estudio equivalente desarrollado con metodología tradicional. En el caso de los cuestionarios en Internet, esto es, compararlos con los mismos cuestionarios administrados en papel. El otro es examinar teóricamente si los resultados siguen una determinada tendencia predictiva. El primer método es un tipo de validez convergente mientras que el segundo se define como validez de constructo.

Leung y Kember (2005 citados en Eiroá et al. 2008) señalan como método aceptado para saber si los sujetos están respondiendo a las mismas preguntas, es decir, al mismo constructo, es contrastar las estructuras de los factores entre los grupos o medidas usando modelos de ecuaciones estructurales. Si los modelos resultantes son equivalentes, puede inferirse que los sujetos evalúan los mismos constructos. Es de esta manera que las demandas de precisión requerida pueden ser distintas si el objetivo es clasificar a una persona (apto vs no apto) o si el objetivo es cuantificar su nivel de rasgo. Razón por la cual se expone que en ocasiones es necesario reflexionar sobre cuándo merece la pena

aplicar una Prueba Psicológica Sistematizada y cuándo no (Wainer, 2000 citado en Eiroá et al. 2008).

3.4.1.2. Enfatizar el olvido de ciertas áreas de aplicación de los test.

Desarrollar nuevos tipos de test es costoso y se corre el riesgo de avanzar casi exclusivamente en contextos aplicados (organizacionales o educativos), donde más recursos económicos se invierten o donde más se precisan soluciones tecnológicas eficientes. Los avances no deben olvidar determinados contextos de medición estrictamente propios de nuestra profesión, como son la evaluación clínica o la evaluación de programas de intervención psicosocial (Campos y cols., 2015).

3.4.1.3. Enfatizar el mal uso de las Pruebas Psicológicas Sistematizadas.

Debido a su inmediata disponibilidad, los nuevos tipos de test pueden aplicarse en contextos inadecuados, por personas no preparadas y realizando inferencias erróneas a partir de las puntuaciones que proporcionan (Aguilar-Espinoza, 2013).

Skitka y Sargis (2005 citado en Eiroá et al. 2008) añaden que esta característica ha incrementado en los últimos años, ello a causa de que cada vez es más sencillo el uso de herramientas para implementar pruebas digitales con mayor rapidez y a un bajo costo. Así mismo, añaden las siguientes limitaciones:

- ➔ Diferencias entre los usuarios de Internet con los no usuarios. Ya que los usuarios de Internet son, en general, más jóvenes, tienen más recursos y mayor nivel cultural que la media de población general.
- ➔ Dificultades para extraer muestras probabilísticas. Exceptuando poblaciones en las que todos los usuarios tienen acceso a un correo electrónico como universidades o empresas. Como veremos más adelante, los cuestionarios en línea son adecuados para el muestreo no probabilístico, como el de conveniencia.

- ➔ Menor probabilidad de participar. La sobrecarga del medio ha causado que los posibles participantes estén sobresaturados de requerimientos para participar en estudios. Existen menos posibilidades de participación que a través de medios tradicionales como el teléfono o en persona.

Un meta análisis realizado por Cook, Heath y Thompson (2000, citado en Eiroá et al. 2008), indica que los factores más importantes para fomentar la tasa de respuesta en cuestionarios en red enviados a través de correo son: un elevado número de contactos, personalizar los correos y hacer contactos preliminares. Multitud de sujetos dejan de responder el cuestionario una vez ya han accedido. Aunque no se encuentra una respuesta clara a este problema, se apunta a cuestiones de diseño. Cuando este es excesivamente complejo, extravagante o incómodo (con barras de navegación que no permiten ver con claridad), con instrucciones confusas y pocas instrucciones de cumplimentación; se pueden producir más abandonos.

La mayoría de investigaciones se han centrado en la posibilidad de que las puntuaciones registradas difieran. Por ejemplo, en psicología de la personalidad se han encontrado diferencias al pasar test de medida de los cinco grandes factores de personalidad (Oswald, Carr y Schimdt, 2001; citados en Eiroá et al. 2008). Pueden atribuirse esas diferencias a que ambas condiciones experimentales son administradas a diferentes muestras que, a pesar de ser grandes, nunca se pueden equiparar. En una investigación al efecto, demuestran la equivalencia del papel e Internet utilizando metodología de validación test-retest (Salgado y Moscoso, 2003; citados en Eiroá et al. 2008).

MÉTODO

Objetivo

Conocer la opinión de satisfacción que los participantes presentan al momento de realizar una prueba psicométrica sistematizada (PPS) basada en una prueba de inteligencia estandarizada de inteligencia en una plataforma de aprendizaje e-learning a través de un Teléfono Celular y una Computadora. Ello con la finalidad de identificar el grado de significancia que los dispositivos inteligentes tienen para con la resolución de una PPS.

Objetivos específicos:

1. *Configurar el test de inteligencia Factor G de Cattell en un L.M.S. Moodle.*
2. *Crear un cuestionario que permita capturar y analizar la opinión de satisfacción que los participantes tengan al llevar a cabo la resolución de una prueba psicométrica sistematizada en Moodle.*
3. *Identificar el grado de significancia que tiene el uso de diversos dispositivos tecnológicos en la resolución de una PPS.*

Hipótesis

De acuerdo a lo descrito por Lucas, Mullins, Luna y McInroy, 1977; Greist y col, 1974; Greist y Klein, 1980 citados en Morales-Ramírez et al (2012) las personas que realizan una prueba psicométrica sistematizada se sienten más confortables cuando tienen que responder a preguntas confidenciales. Por otro lado, Skitka y Sargis (2005 citado en Eiroá et al. 2008) sostienen que las personas que realizan una prueba digital tienden a mostrar insatisfacción con pruebas digitales a causa de que desconocen la manera en que funciona el sistema en el cual se está aplicando la evaluación.

A partir de ello, se estipula las siguientes hipótesis:

- H0: No existe una diferencia significativa entre la media de las calificaciones obtenidas en el Cuestionario de Resolución de una Prueba Psicométrica

Sistematizada (CRPPS) por el grupo de participantes que resolvieron la prueba a través de un Smartphone y los que usaron una computadora.

- H1: Existe una diferencia significativa entre la media de las calificaciones obtenidas en el Cuestionario de Resolución de una Prueba Psicométrica Sistematizada (CRPPS) por el grupo de participantes que resolvieron la prueba a través de un Smartphone y los que usaron una computadora.

Variables

Es de suma importancia conocer la experiencia de satisfacción de los usuarios evaluados con respecto a la resolución de una Prueba Psicométrica Sistematizada (PPS). Por lo tanto, se consideran a las siguientes 3 variables básicas para garantizar una experiencia digital satisfactoria de acuerdo con International *Business Machines Corporation (*2017):

1. *Contenido:* Esta dimensión se refiere a toda aquella información relevante y concreta que se presenta en un sitio web y que permite al usuario llevar a cabo su tarea con mayor facilidad.
2. *Diseño:* Esta dimensión hace alusión a la manera en la cual fue construido el sitio web, es decir, los estilos que predominan y definen a un portal en particular.
3. *Interactividad:* Se comprende a la "Interactividad" como aquel comportamiento que el participante mantiene con un dispositivo inteligente para resolver una prueba psicométrica sistematizada.

Diseño

El presente estudio es de enfoque Cuantitativo de tipo No Experimental (Kerlinger y Lee, 2002). Ello con motivo de que en éste se pretende medir la opinión de satisfacción que los participantes tienen para con una PPS que fue previamente diseñada y configurada en MOODLE.

Los datos obtenidos mediante el cuestionario se emplearon para conocer la opinión individual de los participantes durante la resolución de una prueba psicométrica sistematizada aplicada en una plataforma de aprendizaje *e-learning.*

Muestra

Fue comprendida por un grupo de 10 estudiantes y egresados universitarios que fueron candidatos en proceso de selección y reclutamiento por una empresa especializada en TI. La selección de la muestra fue de tipo No probabilística - autoseleccionado; ello a causa de que es complicado acceder a los candidatos en proceso de reclutamiento y selección por empresas privadas mexicanas dedicadas al desarrollo de nuevas herramientas de tecnología de la información y que empleen sistemas digitales (como MOODLE) para llevar a cabo su proceso de evaluación. Además, resulta difícil contar con participación de sujetos que cumplan con los requisitos mínimos obligatorios que el test "Factor g" estipula para su efectiva aplicación.

Por lo tanto, los criterios de inclusión para este estudio fueron los siguientes:

1. Edad mínima de 18 años.
2. Ser estudiante de bachillerato o Licenciatura.
3. Consentir su participación en la resolución de la PPS.

Los criterios de exclusión fueron los siguientes:

1. Haber presentado el test de inteligencia Factor g en el pasado.
2. Estar bajo los efectos de alguna droga o sustancia nociva.

Una vez completa la muestra de 10 participantes se dividieron en dos grupos con base al tipo de muestreo No probabilístico - propositivo con respecto al tipo de herramienta que éstos emplearían para la resolución de una Prueba Psicométrica Sistematizada. Por lo tanto, el grupo 1 llevó a cabo la resolución del test a través de una computadora. Mientras que el Grupo 2 lo realizó por medio de un smartphone. La selección de los participantes para conformar cada grupo de manera aleatoria simple. Ya que, en realidad no se conocía el total de candidatos a evaluar en el proceso de reclutamiento y selección; por lo cual, se decidió colocar a los mismos en función al momento en que estos se fueran presentando.

Instrumentos

Para medir la experiencia digital de satisfacción de los participantes se construyó el "Cuestionario de Resolución de una Prueba Psicométrica Sistematizada (CRPPS)". (ver anexo 1)

El cuestionario se compone por 15 reactivos que constituyen a las dimensiones de Contenido, Diseño e Interactividad. Los indicadores que comprenden a éstos se definen a estos fueron los siguientes:

- *Contenido:* Como indicadores de ésta se considera al texto e imágenes que componen a la PPS Factor G.
- *Diseño:* Para el caso de esta dimensión se consideran como indicadores a los colores, animación efectos y secuencia lógica que presenta la PPS.
- *Interactividad:* En este caso, se toma como indicadores al momento en el que una persona hace uso de algún elemento del dispositivo inteligente para cumplir la tarea; estos elementos se componen por el teclado (virtual o físico) y mouse (físico o por pantalla táctil) que usan para deslizarse o pulsar alguna opción dentro del sitio web.

Los ítems se diseñaron con base al tipo de respuesta de Escalas Tipo Lickert (Ver Anexo 1), por lo cual cuenta con cinco posibilidades distintas de respuestas tales como:

5. Muy de acuerdo

4. De acuerdo

3. Ni de acuerdo ni en desacuerdo

2. En desacuerdo

1. Totalmente en desacuerdo

Ejemplo:

Tabla 2

Muestra de CRPPS

CONTENIDO					
ÌTEM	**Muy de acuerdo**	**De acuerdo**	**Ni de acuerdo ni en desacuerdo**	**En desacuerdo**	**Muy en desacuerdo**
	5	**4**	**3**	**2**	**1**
1. Las instrucciones fueron cortas y precisas.					
2. Las preguntas fueron claras y concretas.					

Tabla 2: **Ejemplo de presentación de ítems y opciones de respuesta del CRPPS.**

La calificación del cuestionario se realiza de manera aditiva con cada puntaje obtenido por los ítems de las 3 dimensiones. Primero se obtiene un puntaje por cada una de éstas y se ubican en la tabla como se muestra a continuación.

Tabla 3

Dimensiones del CRPPS

Dimensiones	Calificaciones
Contenido	5-25
Diseño	5-25
Interacción	5-25
Total	15-75

Tabla 3: Se muestran los rangos de puntos que puede obtener un participante en cada categoría del CRPPS.

Obteniendo, por lo tanto, un puntaje mínimo probable de "5" y un puntaje máximo de "25" por cada categoría.

Este cuestionario tiene una duración de 5 minutos por participante. Su método de presentación fue, al igual que el test de Factor G, a través de una plataforma de aprendizaje *e-learning*. Éste fue configurado para mostrarse después de haber concluido la PPS a través de un botón con hipervínculo al cuestionario.

Procedimiento

Parte 1. Configuración de la PPS:

Se eligió como base la prueba de inteligencia "Factor G", misma que fue adaptada a un ambiente digital para fines de esta investigación. Ello a causa de que cuenta con atributos favorables para el estudio tales como:

- Adaptada a población mexicana

    - En el año 2000 González et al baremaron esta prueba con un total de 1322 alumnos provenientes de 34 escuelas públicas de Educación Media Superior (Bachillerato) y Educación Superior de la Ciudad de México.

- Contar con una extensión no mayor a 100 reactivos.
    - El test cuenta con 50 reactivos que comprenden a 4 categorías específicas: Series, Clasificación, Matrices, Condiciones. Lo cual reduce la fatiga y aburrimiento del usuario evaluado.
- Rango mínimo de edad de 18 a 25 años.
    - Esto permite llevar a cabo su aplicación en ambientes organizacionales y educativos con adultos.
- Instrucciones sencillas de comprensión.
    - Los ítems están construidos con base a imágenes, lo cual reduce el tiempo y los recursos empleados por el participante para resolver el test.
- Método sencillo de evaluación e interpretación.
    - La valoración de las respuestas capturadas por los participantes es simple. Lo cual permitió configurar la plataforma de aprendizaje Moodle para procesar los datos obtenidos por los candidatos.

Por otro lado, se escogió la plataforma de aprendizaje e-learning MOODLE para la reconstrucción, aplicación y procesador de los datos de la PPS. Ello a raíz de que permite la presentación total de éste, la finalización de la prueba, el cálculo del nivel de habilidad, almacenamiento de los resultados, actualización del banco de ítems y la presentación de un informe escrito al examinado sobre su desempeño (Olea & Ponsoda, 1996).

Parte 2. Creación y configuración del "Cuestionario de Resolución de Prueba Psicométrica Sistematizada (CRPPS)":

Consistió en realizar un análisis del Estado del Arte de la incorporación de las pruebas psicométricas a las nuevas tecnologías de la Información y, con base en ello, formular un cuestionario que contuviera las variables más importantes al momento de construir una prueba psicométrica sistematizada. Para ello se tomó como base el artículo de la empresa International Business Machines (IBM) (2017): "IBM and Automation Anywhere: A new partnership to reinvent business process". El cual establece que los factores más importantes para crear un sitio web eficiente son el Contenido, Diseño e Interactividad. Posteriormente, se crearon los reactivos de acuerdo a las categorías expuestas por IBM y se agregó la escala de respuestas tipo Likert.

Posteriormente, se construyó el cuestionario en la plataforma de aprendizaje e-learning como actividad de "ENCUESTA" y se configuró como actividad secuencial a la PPS con ayuda de un botón que contaba con un hipervínculo hacia ésta.

Finalmente, se entregaron las calificaciones a los participantes inmediatamente después de haber concluido con la PPS y el CRPPS a través de un botón con hipervínculo al "Calificador".

Parte 3. Aplicación de PPS y CRPPS a 10 participantes:

Se crearon 10 usuarios con información aleatoria en la plataforma de aprendizaje e-learning; ello con la finalidad de garantizar el anonimato de los participantes. Así mismo, se configuraron dichos usuarios para que sólo tuvieran acceso a la PPS, el CRPPS y el "Calificador". De esta manera se redujo la probabilidad de que los participantes se distrajeran con los anuncios, actividades, recursos e información propios de la plataforma.

Posteriormente, se seleccionó al grupo de participantes que cumplieran con los requisitos estipulados y se les informó del estudio en cuestión. Así mismo, se les solicitó su aprobación para participar en este estudio. Una vez completa la muestra de 10 participantes, se dividieron en dos grupos: El grupo 1 llevó a cabo la resolución del test a

través de una computadora. Mientras que el Grupo 2 lo realizó por medio de un smartphone.

Una vez creados y organizados los grupos se les prestó una computadora o un smartphone (según fuese el caso) y se les otorgó una breve introducción de lo que consistía la evaluación. Después, se le dieron instrucciones precisas de cómo ingresar a la plataforma y el momento en el cual comenzaría a correr el tiempo de aplicación de la prueba. Razón por la cual se indicó el momento exacto en el que debía iniciar su evaluación.

Finalmente, después de haber obtenido su puntuación en la PPS de inteligencia Factor G y haber resuelto el cuestionario se les agradeció por su amable participación y se les recogieron los instrumentos.

Parte 4. Análisis e interpretación de los datos:

La plataforma de aprendizaje e-learning captura, guarda y valora toda la información de los puntajes obtenidos de los participantes de la PPS y el CRPPS. De esta manera, se obtuvo un reporte en formato XLSX con los puntajes capturados por los participantes con respecto al Cuestionario de Resolución de una Prueba Psicométrica Sistematizada.

Posteriormente, se transfirieron estos datos al programa de procesamiento estadístico SPSS. Con éste se determinó el nivel Alfa de error de significancia de 0.05%, esto en función a lo expuesto por Hernández et al (2014) para con las investigaciones experimentales y cuasi-experimentales de las ciencias sociales y de la salud. Logrando con ello una matriz de datos que permitieran el análisis y descripción de los mismos para determinar la opinión de los participantes con base al valor de las Medidas de Tendencia Central y t de Student para muestras independientes. Esta última se llevó a cabo para conocer el grado de significancia de los puntajes obtenidos por ambos grupos.

RESULTADOS

Análisis de los datos obtenidos en el Cuestionario de Resolución de una Prueba Psicométrica Sistematizada.

En la Figura 1 se muestran los puntajes obtenidos por los participantes en el CRPPS con respecto a la variable de “Contenido”, “Diseño” e “Interactividad”.

Figura 1

Porcentaje de Grupos.

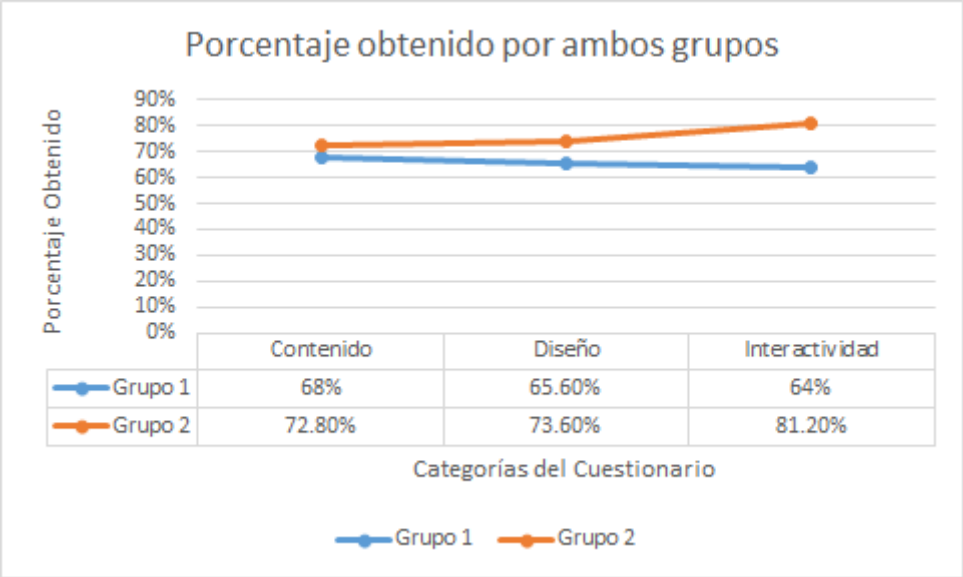

Figura 1: Se muestran los porcentajes de satisfacción obtenidos en el Cuestionario de Resolución de una Prueba Psicométrica Sistematizada por los grupos de evaluación.

El grupo 1 denotó un 68% de satisfacción con las imágenes y los textos empleados en la prueba, mientras que, por su otro lado, el Grupo 2 obtuvo un 73% de satisfacción con éstos.

Con respecto a la variable de “Diseño”, el cual comprendía los colores, animación, efectos y secuencia lógica que presenta la PPS el Grupo 1 estableció estar satisfecho con

éstos en un 66%, mientras que el Grupo 2 estableció un 74% de satisfacción con el diseño de la PPS.

Finalmente, con relación a la variable de “Interactividad”, la cual comprende el uso de algún elemento del dispositivo inteligente para cumplir satisfactoriamente con la PPS; teclado (virtual o físico) y mouse (físico o por pantalla táctil) que se usan para deslizarse o pulsar alguna opción dentro del sitio web; Se obtuvo un porcentaje de satisfacción del 61% por parte del Grupo 1 y un 82% por parte del Grupo 2.

En la Tabla 4 se muestran las Medida de Tendencia Central que obtuvieron ambos grupos:

Tabla 4

Puntajes obtenidos.

	Descriptivos			Estadístico	Error típ.
Calificación	Computadora	Media		49,40	4,354
		Intervalo de confianza para la media al 95%	Límite inferior	37,31	
			Límite superior	61,49	
		Media recortada al 5%		49,06	
		Mediana		46,00	
		Varianza		94,800	
		Desv. típ.		9,737	

		Mínimo		40	
		Máximo		65	
	Smartphone	Media		57,00	5,030
		Intervalo de confianza para la media al 95%	Límite inferior	43,03	
			Límite superior	70,97	
		Media recortada al 5%		56,89	
		Mediana		55,00	
		Varianza		126,500	
		Desv. típ.		11,247	
		Mínimo		43	
		Máximo		73	

Tabla 4: Se muestran las Medidas de Tendencia Central que comprende a los grupos de evaluación que realizaron la PPS a través de una computadora y un smartphone.

En la Tabla 5 se muestra el P-valor de la distribución de VA para ambos grupos obtenido de la prueba de Shapiro Wilk (1965) para grupos =< a 30 sujetos.

Tabla 5.

Pruebas de normalidad.

Pruebas de normalidad

	¿Cuál fue el el instrumento que usó para resolver el test?	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Estadístico	gl	Sig.	Estadístico	gl	Sig.
Calificación	Computadora	,237	5	,200*	,907	5	,451
	Smartphone	,171	5	,200*	,987	5	,970

*. Este es un límite inferior de la significación verdadera.
a. Corrección de la significación de Lilliefors

Tabla 5: Se muestran los valores de distribución de las VA de los estadísticos

En la tabla 6 se resumen los valores de la distribución de las VA

Tabla 6

Normalización de puntajes.

NORMALIDAD (Calificaciones)	
P-Valor (Smartphone) = 0.451	**Alpha=0.05**
P-Valor (Computadora) = 0.970	**Alpha=0.05**

Tabla 6: La distribución de la variable aleatoria de P-Valor para ambos grupos se comporta normalmente.

Por lo tanto, se acepta H0 y se corrobora que se cumple el criterio de normalidad para ejecutar el estadístico t de Student.

En la tabla 7 se muestra el resultado de este cálculo del estadístico de Levene (1960) para evaluar la igualdad de las varianzas para una variable calculada para dos o más grupos.

Tabla 7

Prueba Levene

		Prueba de Levene para la igualdad de varianzas	
		F	**Sig.**
Calificación	**Se han asumido varianzas iguales**	**,094**	**,767**
	No se han asumido varianzas iguales		

Tabla 7. Se muestra la estimación de las varianzas de los grupos que resolvieron la PPS a través de una computadora y un smartphone.

De esta manera la Igualdad de varianza (Prueba de LEVENE) se puede resumir en el siguiente cuadro comparativo:

IGUALDAD DE VARIANZA		
P-Valor = 0.767	>	Alpha = 0.05

Por lo tanto, se acepta H0 y se comprueba que las varianzas de ambos casos son iguales.

La tabla 8 muestra el resultado del estadístico t de student obtenido para muestras independientes.

Tabla 8

Prueba t de Student para muestras independientes

		Prueba T para la igualdad de medias						
		t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
							Inferior	Superior
Calificación	Se han asumido varianzas iguales	-1,142	8	,286	-7,600	6,653	-22,941	7,741
	No se han asumido varianzas iguales	-1,142	7,839	,287	-7,600	6,653	-22,996	7,796

Tabla 8: Se observa el resultado de la estimación del estadístico t de student para los participantes del grupo 1 y 2 del estudio.

Se resume el cálculo obtenido de t de student en el siguiente cuadro:

Calcula nivel de significancia t de Student		
P-Valor = 0.286	>	Alpha = 0.05

Por lo tanto, se acepta H0 y se establece que la diferencia entre la media de calificaciones obtenidas por el grupo de participantes que resolvieron la PPS a través de un smartphone y los que usaron una computadora no es estadísticamente significativa.

DISCUSIÓN

Existe una gran diversidad de métodos y técnicas para realizar un proceso de selección y reclutamiento de personal eficaz. Sin embargo, muchos de esos métodos se ven comprometidos en ocasiones a causa de la manera en que éstos son ejecutados por el personal administrativo de las organizaciones asignados a esta tarea, entrevistas, exámenes técnicos, exámenes de conocimientos y pruebas psicométricas. Ello ocurre a causa de la creciente demanda de las industrias por conseguir personal competente y eficaz que impulse su producción y permita cumplir con los objetivos organizacionales planteados por la misma (Aamodt, 2010)

No obstante, a causa de esta gran demanda se ha suscitado un crecimiento exponencial hacia el uso de las nuevas tecnologías de la información que impulsen el crecimiento del Factor Humano en las organizaciones. Mismas que permite a las organizaciones ejecutar procesos de administración del personal, educación y evaluaciones (Aguilar-Espinoza, 2013). Con respecto a este último, las investigaciones que se han realizado al respecto giran en torno al uso de pruebas psicométricas sistematizadas (digitales) que permitan a los evaluadores procesar los datos obtenidos de todos los participantes implicados en cuestión de segundos (Priále, 1983, citado en Aguilar-Espinoza, 2013); poner a prueba nuevos métodos de construcción y análisis de datos de pruebas psicométricas aplicadas a personal organizacional (Escurra, Delgado y Aparcana, 1993) y sistemas digitales (Software) que den pie a configurar y aplicar pruebas psicométricas y capturar, procesar e interpretar los datos obtenidos de manera automatizada (Mendenhall, 2017, Tirado et al, 2016, Campos y cols., 2015,González y Hernández, 2013, Bruning, Schraw y Norby, 2012, Molinar et al., 2012, Morales-Ramírez et al, 2012, Pérez, 2008, Sierra–Matamoros et al, 2007, Abad et al., 2006, Ponsoda et al., 2004, The Mathworks Inc., 2001).

En esta investigación se tomó como objeto de estudio la experiencia de los participantes evaluados con respecto a la resolución de una prueba psicométrica sistematizada. Es decir, se consideró la percepción que los participantes tuvieron para con

la Prueba Psicométrica Sistematizada (PPS) como factor principal. Ello a causa de que no existen investigaciones previas que aborden factores individuales que puedan intervenir en los datos obtenidos por un test sistematizado. Lo cual, dentro de la Teoría Clásica de los Test es denominado como "Errores de Medición" (Nunnally, 1987 y Magnusson, 1980). Mismos que establece el método de aplicación con el cual se ejecuta la evaluación puede favorecer o empobrecer los puntajes obtenidos por los evaluados (Brennan, 2001 y Cronbach, et al., 1972, citados en Martínez, et al, 2014). Es por ello que, en el esfuerzo por postular una propuesta de instrumento que contribuya al desarrollo de esta línea de trabajo se construyó el Cuestionario de Resolución de una Prueba Psicométrica Sistematizada (CRPSS), el cual fue orientado a capturar el nivel de percepción de satisfacción que los usuarios tuvieron al momento de resolver la PPS. Éste se conformó por las variables que, de acuerdo con *International Business Machines Corporation* (2017) indica como factores esenciales para la construcción de un sitio web exitoso y atractivo para el público: 1) Contenido, 2) Diseño y 3) Interactividad.

Con base en los resultados obtenidos se puede inferir que, si bien que el grado de significancia de ambas pruebas es nula ($t=0.286 > 0.05$), es evidente que el puntaje obtenido a través de distintos medios tecnológicos es notorio (Grupo 1 $=\bar{X}=49.40$, Grupo 2= $\bar{X}=57$), aún con un grupo muestral de tan sólo 5 individuos por grupo. Sin embargo, al no haber sido aleatorizados los datos no se puede asegurar que las diferencias sean a causa de los dispositivos.

No obstante, la plataforma de aprendizaje Moodle no cuenta con los suficientes recursos y/o facilidades como para construir una interfaz más sofisticada e intuitiva para los participantes; lo cual pudo afectar de cierta forma los puntajes capturados por los usuarios. Así mismo, también se sugiere realizar un instrumento válido y confiable que valore la percepción de los participantes sobre la resolución de una prueba psicométrica sistematizada. También, se insta a que se replique este estudio cambiando el portal por el cual se aplica la PPS como bien pueden ser los Softwares As A Services (S.A.A.S.) o Content Management System (C.M.S.).

Por otro lado, cabe señalar que los participantes considerados en este estudio estuvieron bajo un proceso de reclutamiento y selección por parte de una empresa privada dedicada al área de servicio con base a TI, lo cual pudo alterar las calificaciones que éstos

tuvieron para con la PPS. También, de acuerdo con Hernández, Fernández y Baptista (2014), al no ser una muestra representativa y no contar con el supuesto de normalidad, los datos no pueden ser generalizables, por lo que el poder explicativo de instrumento no sería necesariamente útil, para explicar cualquier otro conjunto de sujetos.

Finalmente, es de suma importancia señalar que este estudio sólo retomó a jóvenes estudiantes y/o pasantes de la carrera universitaria de la Universidad Nacional Autónoma de México (UNAM) de edad entre 23 y 26 años; respetando de esta manera los criterios con los cuales fue baremado el Test de inteligencia Factor G por Gonzáles et al en el año 2000. Razón por la cual se insta a que en futuras investigaciones se tome a una muestra con características distintas sin violar los criterios de inclusión de la prueba como bien pueden ser PPS neuropsicológicas aplicados a través de instrumentos computacionales a una población infantil o adultos mayores.

CONCLUSIÓN

Los resultados obtenidos en esta investigación permiten observar que, en efecto, existen diferencias notorias en las puntuaciones obtenidas en el Cuestionario de Resolución de una Prueba Psicométrica Sistematizada (CRPPS) por los candidatos que realizaron la PPS a través de un ordenador PC y de los que la realizaron a través de un smartphone. Denotando de esta manera que la estructura con la cual fue construido el test sistematizado afecta a la percepción satisfactoria que los participantes demuestran para con el proceso de evaluación que se está ejecutando. Lo cual, de acuerdo con Catalán y González (2010), es una variable importante que podría afectar gravemente a las puntuaciones recabadas por el instrumento; independientemente de si éste cuenta con niveles elevados de validez y confiabilidad en aplicación a lápiz y papel.

Por otro lado, existe evidencia de que las nuevas tecnologías de la información son una herramienta poderosa para impulsar el uso de test psicométricos de calidad (Tirado, et al., 2016, Campos, et al. 2015, Liporace y Solano, 2015, Aguilar-Espinoza, 2013, González, E. y Hernández, 2013, Molinar et al., 2012, Morales-Ramirez et al., 2012, Olea et al., 2010, Sierra-Matamoros et al., 2007). Sin embargo, no hay investigaciones que aborden el grado de afectación que éstas pueden surgir a raíz de la percepción favorable o desfavorable, satisfactoria o insatisfactoria que los participantes a ser evaluados obtengan en un proceso de reclutamiento y selección, capacitación y adiestramiento del personal, evaluación de desempeño laboral, procesos de consultoría y gestión de proyectos, entre otros.

Es por tanto que este estudio permite a los futuros investigadores reflexionar sobre los métodos, instrumentos y habilidades individuales que los participantes sujetos a ser evaluados y los evaluadores deben tomar en consideración para no ver comprometidos sus resultados al momento de incursionar en el área organizacional y encontrarse en una situación en la cual sea esencial ejecutar pruebas psicométricas sistematizadas.

También es de suma importancia mencionar que si bien es un hecho que la tecnología puede impulsar el crecimiento de la psicología de manera exponencial también

puede perjudicarla gravemente si los futuros profesionistas no aprenden a gestionar este tipo de procesos organizacionales propios de la ciencia a través de nuevas herramientas Tecnológicas de la Información (TI). Ello a raíz de que una de las grandes limitantes que surgieron en este estudio fue la selección y configuración de la plataforma de aprendizaje e-learning para lograr la adaptación exitosa del "Factor g". Misma que cumplió su cometido gracias al trabajo interdisciplinario de ingenieros en programación y desarrollo y Bases de datos, mismos que se ofrecieron voluntariamente a colaborar con el estudio.

Finalmente, se recomienda a los futuros investigadores continuar con esta línea de investigación ya que actualmente *"Los clientes y consumidores ya han trasladado su presencia y la conversación al entorno digital, y las empresas tienen que escucharles e interactuar con ellos."* (Lopez, 2018). Y depende enteramente de los psicólogos adaptar las herramientas de evaluación a los medios digitales para impulsar el desarrollo de la ciencia en la industria y ofrecer soluciones de cambio con ayuda de las nuevas tecnologías de la información sin olvidar el factor más importante de la evaluación; El individuo. Resumiendo, con esto, lo mencionado por B.F. Skinner sobre las nuevas tecnologías: "El problema real no es si las máquinas piensan, sino si lo hacen los hombres. (Skinner, 1978).

REFERENCIAS

Aamodt, M. (2010). Psicología Industrial/Organizacional. México: Cengage Learning.

Abad, F., Garrido, J., Olea, J. y Ponsoda, V. (2006). Introducción a la psicometría. España: *Universidad Autónoma de Madrid.*

Aguilar-Espinoza, J. (2013). Pruebas Psicológicas Sistematizadas. *Asistencia Electrónica de Información Psicológica*.

http://www.academia.edu/8113908/Pruebas_Psicol%C3%B3gicas_Sistematizadas

Anastasi, A. (1988). Psychological Testing. New York: MacMillan Publishing Company.

Aragón, L. (2004). Fundamentos Psicométricos en la evaluación psicológica. *Revista Electrónica de Psicología Iztacala*. 7(4):23-43. http://www.iztacala.unam.mx/carreras/psicologia/psiclin/vol7num4/Art3-2005-1.pdf

Argibay, J. (2006). Técnicas psicométricas. Cuestiones de validez y confiabilidad. Subjetividad y procesos cognitivos. (8):15-33.

http://www.redalyc.org/pdf/3396/339630247002.pdf

Barbero, M., Vila, E. y Holgado, F. (2010). Psicometría. España: Sanz y Torres.

Birkett, H. (1993). Lo profundo de la personalidad, aplicación del 16FP. México: Manual Moderno.

Blum, M. y Naylor, J. (1985). Psicología industrial. México: Trillas.

Bruning, R., Schraw, G. y Norby, M. (2012) Psicología cognitiva y de la instrucción. Estados Unidos de América: Pearson.

Campos, D., Quero, S., Bretón, J., Riera, A., Mira, A., Tortella, M. y Botella, C. (2015). Correlation between psychological evaluation through the internet and traditional evaluation applied by the therapist for flying phobia. *Tesis Psicológica*. 10(2): 52-67. Recuperado en http://www.redalyc.org/pdf/1390/139046451004.pdf

Catillo, J. y Folino, J. (2009). VALIDEZ CONCURRENTE Y PREDICTIVA DE LA ESCALA DE CRIBADO DE ESTILO DE VIDA DELICTIVO REVISADA -ECEViD R. Revista de la Facultad de Medicina. 57(4):295-303.

Díaz, C., Botanero C. y Cobo, B. (2003). Fiabilidad y generalizabilidad. Aplicaciones en la evaluación educativa. *Números*. 54:3-21. Recuperado en: http://www.sinewton.org/numeros/numeros/54/Articulo01.pdf

Dunnette, M. y Kirchner, W. (1989). Psicología Industrial. México: Trillas.

Eiroá, F., Fernández, I. y Pérez, P. (2008). Cuestionarios psicológicos e investigación en Internet: Una revisión de la literatura. *Anales de psicología*. 24(1): 150-157. Recuperado en: http://www.redalyc.org/html/167/16724119/

Fernández, M., Noelia, A. y Antonio, M. (2009). Curso básico de psicometría. México: Lugar Editorial.

Gempp, R. (2006). El error estándar de medida y la verdadera puntuación de los test psicológicos: Algunas recomendaciones prácticas. *Terapia Psicológica*. 24(2): 117-130. Recuperado en: https://www.researchgate.net/publication/26493181_El_error_estandar_de_medida_y_la_puntuacion_verdadera_de_los_tests_psicologicos_Algunas_recomendaciones_practicas

Gómez-Benito, J., Hidalgo, M. y Guilera, G. (2010). El sesgo de los instrumentos de medición. Test justos. *Papeles del Psicólogo*. 31(1): 75-84.

González, E. y Hernández, D. (2013). Batería psicológica sistematizada para estudiantes universitarios encaminado a una educación integral. *Revista Electrónica de Psicología Iztacala*. 16(1): 269-287. Recuperado en:

http://www.iztacala.unam.mx/carreras/psicologia/psiclin/vol16num1/Vol16No1Art16.pdf

González, F. (2007). Instrumentos de Evaluación Psicológica. Cuba: Ciencias Médicas.

González, M., Aragón, L. y Silva, A. (2000). Baremación del test de inteligencia factor «G» de Cattell, en la zona metropolitana de la ciudad de México. *Psicothema.* 12(2):275-278. http://www.psicothema.com/pdf/564.pdf.

Guth, A. (2004). Reclutamiento, selección e integración de Recursos Humanos. México: Trillas.

Hoste, R. (1981). How Valid are School Examination? An explotation into Content Validity. *British Journal of Educational Psychology*. ? (51):10-22. Recuperado en: http://onlinelibrary.wiley.com/doi/10.1111/j.2044-8279.1981.tb02450.x/abstract

Kim, S. y Wilson, M. (2009). A comparative Analysis of the Ratings in performance assessment using generalizability theory and the many-facet rasch model. *Journal of Applied Measurement*. 10(4):408-423. Recuperado en: https://www.ncbi.nlm.nih.gov/pubmed/19934528

Landy, F. y Conte, J. (2005). Psicología industrial: Introducción a la psicología industrial y organizacional. México: McGraw-Hill.

Lanyon, R. y Goodstein, L. (1977). Evaluación de la personalidad. México: Manual Moderno.

Larsen, R. y Buss, D. (2005). Psicología de la personalidad. México: McGraw-Hill.

Liporace, M. y Solano, A. (2015). Evaluación de la personalidad normal y sus trastornos. Argentina: Lugar Editorial.

Magnusson, D. (1977). Teoría de los test. México: Trillas.

Martínez, M, Hernández, M. y Hernández, M. (2014). Psicometría. Argentina: Alianza.

Matesanz, A. (1998). Evaluación estructurada de la personalidad. España: Pirámide.

Mendenhall, W., Beaver, R. y Beaver, B. (2017). Probabilidad y estadística para las ciencias sociales del comportamiento y la salud. México: CENGAGE LEARNING.

Molinar, J., Escoto, M., García, R. y Bautista E. (2012). Evaluación Computarizada de Pruebas Psicológicas Mediante el Procesamiento Digital de Imágenes. *Enseñanza e Investigación en Psicología.* 17(2): 415-426. Recuperado en http://www.redalyc.org/pdf/292/29224159006.pdf

Morales-Ramírez, A., Escoto, M., García-Lozano, R., Molinar-Solís, J. e Hidalgo-Cortés, C. (2012). Sistema para la aplicación de pruebas psicológicas vía web. *Acta Universitaria.* 22(3): 5-13. Recuperado en http://www.acuedi.org/ddata/1689.pdf

Moreno, B. (2007). Psicología de la personalidad. España: Thomson.

Muñiz, J. (2010). Las teorías de los test: teoría clásica y teoría de respuesta a los ítems. *Papeles del psicólogo*. 31(1):57-66. http://www.redalyc.org/pdf/778/77812441006.pdf

Nunnally, J. (1987). Teoría Psicométrica. México: Trillas.

Olea, J., Abad, F.J y Barrada, J.R. (2010). Test informatizados y otros nuevos tipos de test. *Papeles del Psicólogo*, 31(1), 97-107. Recuperado en: http://www.redalyc.org/pdf/778/Resumenes/Resumen_77812441010_1.pdf

Prieto, G. y Delgado, A. (2010). Fiabilidad y validez. Papeles del Psicólogo, 31(1), 67-74

Sierra–Matamoros, F., Valdelamar–Jiménez, J., Hernández–Tamayo, F. y Sarmiento–García, L. (2007). Test adaptativos informatizados. *Avances en*

Medición. 5: 157-162. Recuperado en http://www.humanas.unal.edu.co/psicometria/files/9513/7036/5476/Test_Adaptativos_Informatizados.pdf

Spector, P. (2002). Psicología industrial y organizacional. México: Manual Moderno.

Terán, L. (2017). Inicios de la psicología. Recuperado en: https://datospdf.com/download/inicios-de-la-psicologia-_5a4b70b9b7d7bcb74faf87e4_pdf

Tirado, F., Backhoff, E. y Larrazolo, N. (2016). La revolución digital y la evaluación: un nuevo paradigma. *Perfiles Educativos*. 38(152):182-201. Recuperado en: http://www.redalyc.org/pdf/132/Resumenes/Abstract_13244824011_2.pdf

Usabiaga, O., Castellano, J., Blanco-Villaseñor, Á. y Casamichana, D. (2013). La teoría de la generalizabilidad en las primeras fases del método observacional aplicado en el ámbito de la iniciación deportiva: calidad del dato y estimación de la muestra. *Revista Psicológica del Deporte*. 22(1):103-109. Recuperado en:

http://www.redalyc.org/pdf/2351/235127552014.pdf

Zúñiga-Brenes, M. y Montero-Rojas, E. (2007). Teoría G: un futuro paradigma para el análisis de pruebas psicométricas. *Actualidades en Psicología*. 21:117-144. Recuperado en:

http://pepsic.bvsalud.org/pdf/apsi/v21n108/v21n108a06.pdf

ANEXOS

Anexo 1

Cuestionario de Resolución de una Prueba Psicométrica Sistematizada

Introducción

A continuación, se presenta un cuestionario que tiene por objetivo describir la experiencia que una persona presenta al momento de resolver una prueba psicométrica sistematizada. Con estos datos pretendemos obtener información clara y objetiva sobre la opinión de los participantes con respecto a las pruebas psicológicas digitales; las cuales se implementan actualmente de manera masiva en el ámbito organizacional para los procesos de reclutamiento, selección y capacitación de personal a través de las nuevas tecnologías de la información.

Es de suma importancia mencionar que toda la información que usted nos proporcione a través de este cuestionario totalmente anónimos, confidenciales y con fines exclusivos de investigación. Por lo cual, le pedimos de la manera más atenta que conteste el instrumento con suma franqueza.

Finalmente, le agradecemos enteramente su participación y el tiempo que destino para la resolución de la prueba y el cuestionario de resolución de una prueba sistematizada.

CONTENIDO					
ÌTEM	**Muy de acuerdo**	**De acuerdo**	**No estoy de acuerdo ni en desacuerdo**	**En desacuerdo**	**Muy en desacuerdo**
	5	**4**	**3**	**2**	**1**

1. Las instrucciones fueron cortas y precisas.					
2. Las preguntas fueron claras y concretas.					
3. Las imágenes tenían un tamaño adecuado para la prueba.					
4. La calidad de las imágenes fue oportuno para la prueba.					
5. Usar preguntas e imágenes fue pertinente para resolver el test.					
DISEÑO					
ÌTEM	**Estoy Muy de acuerdo**	**Estoy De acuerdo**	**No estoy de acuerdo ni en desacuerdo**	**Estoy en desacuerdo**	**Estoy muy en desacuerdo**
6. Los colores de las imágenes fueron efectivos para resolver el test.					
7. El “efecto de Arrastrar y Soltar” fue oportuno para resolver el test.					

8. La secuencia de las preguntas fue adecuado para la resolución del test.					
9. Las animaciones de los avisos de la prueba fueron efectivos.					
10. La secuencia con la que se presentó el test, el cuestionario y los resultados fue oportuna.					
INTERACCIÓN					
ÌTEM	**Estoy Muy de acuerdo**	**Estoy De acuerdo**	**No estoy de acuerdo ni en desacuerdo**	**Estoy en desacuerdo**	**Estoy muy en desacuerdo**
11. No necesité de más de 3 “clics” para llegar a la prueba					
12. Los elementos de mí dispositivos fueron suficientes para resolver la prueba					
13. Pude desplazarme por todo el test sin necesidad de conocer todo el sitio web					
14. Con ayuda del dispositivo inteligente realicé el test en un par de minutos.					

15. Pude obtener fácilmente los resultados de mi prueba en mi dispositivo.					