# REQUISITOS DE ESTUDIOS EMPÍRICOS PARA CONFORMAR UN CUERPO DE CONOCIMIENTOS SÓLIDO: APLICACIÓN A TÉCNICAS DE TESTING

Ana M. Moreno, Sira Vegas

Facultad de Informática – Universidad Politécnica de Madrid
ammoreno@fi.upm.es, svegas@fi.upm.es

A lo largo de 25 años se han venido realizando distintos estudios empíricos sobre técnicas de testing. La cantidad de estudios realizados podría llevar a pensar que en la actualidad disponemos de un cuerpo de conocimientos empíricos sobre técnicas de pruebas de tamaño considerable. Sin embargo, los estudios empíricos realizados en este área poseen carencias importantes que hacen que el cuerpo de conocimientos que podrían proporcionar esté lejos de ser sólido y fiable. En este articulo, utilizamos los resultados de una revisión de estudios empíricos sobre técnicas de testing para identificar y proponer líneas de solución a los problemas encontrados. Las soluciones propuestas están relacionadas fundamentalmente con dos áreas: (1) la rigurosidad del planteamiento del estudio empírico y del análisis de los datos recogidos, y (2) la necesidad de una serie de acuerdos a nivel de comunidad para coordinar la investigación empírica y asegurar que los estudios empíricos que se realizan se complementan y ratifican unos a otros.

**Palabras clave:** Experimentos, Cuerpo de conocimientos empírico, Base de experiencias, técnicas de Testing

## 1. Introducción

La comunidad científica de desarrollo de software ha generado a lo largo de los últimos 25 años diversos estudios empíricos que tratan de analizar las características de las técnicas de testing con objeto de determinar los beneficios de unas frente a otras, así como identificar las mejores condiciones de aplicación de dichas técnicas. Este tipo de información, generada a partir de distintos estudios empíricos consistentes y complementarios, permitiría constituir un cuerpo de conocimientos empíricos.

En [10] y [14] las autoras han realizado una revisión exhaustiva de los estudios empíricos publicados sobre técnicas de testing con el fin de extraer las unidades de conocimiento que podrían conformar el cuerpo empírico sobre testing. La conclusión de este trabajo fue que los resultados de los estudios empíricos realizados sobre técnicas de testing no poseen el suficiente nivel de madurez como para dar lugar a un cuerpo de conocimiento sólido sobre testing. Esta falta de madurez en la contrastación empírica de conocimientos, no es exclusivo del área de testing, sino que es un problema generalizado a otras áreas de la IS [20].

Este artículo pretende dar un paso más en este sentido, analizando con detalle cuáles son los motivos o problemas que causan la poca madurez empírica del cuerpo de conocimiento actual en técnicas de testing y proponiendo líneas de solución para estos problemas.

Con este fin, la sección 2 describe las características revisadas sobre los estudios empíricos de [10] para determinar su grado de madurez. La sección 3 presenta los problemas detectados en estos estudios, así como posibles guías que ayudarían a la creación de un cuerpo de conocimiento empírico. Finalmente la sección 4 presenta las conclusiones obtenidas.

## 2. Caracterización de la Madurez de Estudios Empíricos

Los estudios empíricos han de realizarse con el rigor necesario que garantice que los resultados serán válidos. Para poder evaluar la fiabilidad de los resultados de los estudios empíricos existentes sobre técnicas de testing, hemos analizado los siguientes criterios:

- *Rigurosidad en el planteamiento del estudio*. Existen distintas circunstancias que pueden hacer que el planteamiento de un estudio empírico sea inadecuado y, por lo tanto, sus resultados no sean fiables. Estas circunstancias son lo que se conoce como amenazas a la validez [8] [23], y están relacionadas, por ejemplo, con el orden de aplicación de las técnicas a comparar, la selección de los sujetos que intervienen en los estudios, etc. Así, en los estudios revisados se ha analizado cómo han sido tratadas las amenazas a la validez correspondientes.
- *Rigurosidad en el análisis de los datos obtenidos del estudio.* Es necesario que los datos obtenidos de los estudios empíricos sean analizados con técnicas que permitan extraer de ellos resultados objetivos. Por ello, se ha prestado especial atención al grado de confiabilidad de los resultados arrojados por las técnicas de análisis utilizadas en los experimentos.
- *Establecimiento de conclusiones que trasciendan el mero análisis de los datos obtenidos*. La generación de conocimiento empírico se realiza a partir de la definición de unos objetivos generales que se van contrastando a través de distintos estudios empíricos. Por ello, también hemos comprobado si los estudios se han limitado meramente a describir los resultados arrojados por las técnicas de análisis, o si por el contrario, han elaborado conclusiones de más alto nivel relacionadas con los objetivos generales a los que podría contribuir dicho estudio empírico permitiendo así, generar un item de conocimiento y facilitando replicaciones y estudios futuros consecutivos.

Los estudios empíricos han de arrojar resultados que sean significativos y útiles en situaciones reales. Esto implica que es importante que haya una correspondencia entre las condiciones en las que se realiza el estudio empírico y las condiciones en las que las técnicas estudiadas se aplican en la práctica, en caso contrario, los resultados de dicho estudio serían poco útiles. Para comprobar esta concordancia se han analizado las siguientes características

- *Utilización de programas y faltas representativas.* Se ha observado si los estudios analizados han utilizado programas y faltas representativas de las que ocurren en la realidad. Para esto, se ha tenido en cuenta tanto el tamaño y características de los programas, como el número y naturaleza de las faltas.
- *Variables respuesta de interés*. En este sentido, se ha tenido en cuenta la utilidad de las métricas usadas para recoger los datos del estudio empírico. El grado de utilidad se ha establecido en función de cuánto de útil sería esa métrica para los ingenieros software en su práctica profesional.
- *Uso realista de la técnica.* También se ha querido analizar hasta qué punto la ejecución del estudio es representativa de la realidad. Así que en este caso se ha comprobado en qué medida los estudios empíricos revisados utilizan las técnicas de testing de una manera semejante a como se aplican en la realidad.

Un cuerpo de conocimiento en cualquier disciplina no se construye con un único estudio empírico, ni con estudios empíricos aislados. Por este motivo se ha estudiado cómo los distintos estudios revisados facilitan o permiten la complementación de sus resultados con otros estudios que se realicen a partir de ellos, ya sea como replicaciones tal cual, o como ampliaciones de los mismos. Este aspecto incluye las siguientes características:

- *Se proporcional detalle suficiente que permite la replicación*. Esto es, se ha comprobado en qué medida los estudios revisados definen todos los elementos involucrados en el diseño y ejecución del estudio de manera que se pueda replicar.
- *Se encadenan los estudios empíricos de modo que unos recojan y profundicen en las conclusiones de los otros*. Se ha estudiado hasta qué punto existe relación entre los distintos estudios empíricos que conforman cada grupo y cómo se complementan o encajan unos con otros.
- *Avance metodológico en la secuencia de contrastación empírica* de forma que los resultados de los distintos estudios sean progresivamente generalizables. Con este fin, se ha analizado tanto el número de replicaciones como los tipos de estudios empíricos realizados y su secuencia.

Nótese que estos aspectos son suficientemente generales como para poder aplicarse en distintos dominios, esto es, sobre estudios que traten distintos tipos de técnicas, y no exclusivamente técnicas de testing como es el caso de nuestro trabajo.

Volviendo a nuestro trabajo, la Tabla 1 muestra el grado en el que los experimentos analizados cumplen las características mencionadas. Los dos primeros grupos de estudios comparan técnicas de testing pertenecientes a una misma familia, mientras que los tres últimos comparan técnicas de testing pertenecientes a distintas familias. El color blanco en una celda indica que la característica evaluada no se satisface en el estudio correspondiente; el tono gris indica que la característica se satisface en un grado medio, y el tono negro indica que la característica se satisface adecuadamente. Para analizar en detalle cada estudio de la Tabla 1, se refiere a los lectores a [10].

Contaríamos con un cuerpo de conocimientos empíricos sólido sobre técnicas de testing si la tabla fuera completamente negra. A medida que los colores se hacen más claros, indica una mayor cantidad de limitaciones, y, por tanto, pueden considerarse lagunas en el conocimiento empírico sobre técnicas de testing.

## 3. Guías para Madurar el Cuerpo de Conocimientos Empíricos

A partir de las características identificadas en la sección anterior, se ha elaborado una serie de guías que pretenden contribuir a la obtención de un cuerpo de conocimiento empírico sólido sobre técnicas de testing. Estas guías se describen en detalle a continuación.

- **Planteamiento Riguroso**: La realización de un experimento debe contemplar detalladamente todos aquellos elementos que puedan hacer que bien los resultados no sean válidos o no sean generalizables. Estos elementos se denominan amenazas a la validez. Tratar adecuadamente las amenazas a la validez implica definir detalladamente los factores, parámetros variables respuesta del experimento y el diseño del mismo, de manera que dichas amenazas sean minimizadas. Tal como se puede observar en la Tabla 1, salvo en el último grupo de estudios empíricos, en el resto existen problemas de amenazas a la validez. Un estudio detallado de las posibles amenazas a la validez de un experimento y cómo gestionarlas se puede encontrar en [24], distinguiéndose entre amenazas internas y externas según si lo que provocan son la invalidez de los resultados obtenidos o su falta de generalidad.

- **Análisis Riguroso de los Datos:** Otra de las características fundamentales que debe tener un experimento es la realización de un análisis riguroso de los datos obtenidos en el mismo. En general, puede observarse, según la Tabla 1, que aproximadamente la mitad de los estudios no respaldan sus conclusiones con análisis rigurosos. En su lugar realizan un análisis grosso modo basado en la interpretación cualitativa de los datos por parte de los autores.

  La forma de conseguir un análisis fiable de los datos obtenidos en un estudio empírico es mediante la utilización de técnicas formales de análisis de datos. Las técnicas que tradicionalmente se aplican para realizar estos análisis suelen ser técnicas estadísticas. Sin embargo, también es posible la utilización de técnicas de análisis cualitativo o incluso basadas en razonamiento fuzzy [23] [8] [13].

- **Establecer Conclusiones de Alto Nivel:** La Tabla 1 muestra que tan solo [22] [1] establecen conclusiones más allá del mero análisis de los datos extraídos. [11] [25] sufren sólo parcialmente el problema y el resto de estudios no realizan conclusiones de alto nivel que permitan realizar recomendaciones que otros estudios empíricos retomen.

  Para alcanzar conocimiento empírico es necesario definir objetivos de investigación globales a los que suceden estudios empíricos que van generando el conocimiento deseado [2]. Este conocimiento genera a su vez nuevas metas u objetivos de investigación, para comenzar de nuevo el ciclo. Esto implica que un estudio empírico no debe limitarse al mero análisis de los datos que obtiene, sino que debe, a partir de dichos datos, establecer conclusiones de alto nivel que permitan generar nuevas hipótesis para refinar el conocimiento obtenido.

**Tabla 1. Madurez de los estudios empíricos por familia**

| CARACTERÍSTICA | TÉCNICAS DE FLUJO DE DATOS | | TÉCNCIAS DE MUTACIÓN | | | FLUJO DE CONTROL VS FLUJO DE DATOS | | | MUTACIÓN VS FLUJO DE DATOS | | FUNCIONAL VS FLUJO DE CONTROL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Weyuker [22] | Bieman & Schultz [3] | Offut & Lee [16] | Offut *et al.* [15] | Wong & Mathur [24] | Frankl & Weiss [6] | Hutchins *et al.* [7] | Frankl & Iakounenko [4] | Frankl *et al.* [5] | Wong & Mathur [24] | Basili & Selby [1] | Kamsties & Lott [11] | Wood *et al.* [25] |
| Rigurosidad en el planteamiento del estudio | | | | | | | | | | | | | |
| Rigurosidad en el análisis de los datos obtenidos | | | | | | | | | | | | | |
| Establecimiento de conclusiones que trasciendan el mero análisis | | | | | | | | | | | | | |
| Uso de: programas y/o faltas representativas de la realidad | N/A | N/A | | | | | | | | | | | |
| Las variables respuesta son de interés para practitioners | | | | | | | | | | | | | |
| Uso realista de la técnica | | | | | | | | | | | | | |
| Se proporcionan suficientes detalles para replicación | | | | | | | | | | | | | |
| Encadenamiento de experimentos | | | | | | | | | | | | | |
| Avance metodológico en la secuencia de experimentación | | | | | | | | | | | | | |

El hecho de que un experimentador se detenga en el análisis de los datos obtenidos, tal como ocurre generalmente en los estudios revisados en la Tabla 1, hace que el ciclo anteriormente mencionado se corte, evitando de este modo que, ya sea el propio investigador, ya sean otros grupos, sigan profundizando en ese punto de conocimiento para madurarlo empíricamente.

- **Utilizar Parámetros y Factores Representativos de la Realidad:** En el caso concreto de las técnicas de testing los programas utilizados durante los estudios empíricos, así como las faltas que éstos contienen deben ser lo suficientemente representativas de la realidad como para que los resultados del estudio sean significativos.

  En la Tabla 1, se puede observar que ninguno de los estudios trata totalmente este problema. Bien es cierto que los experimentos de laboratorio controlados (los más abundantes en la Tabla 1), por definición, son de menor dimensión que otro tipo de estudios empíricos, lo que implica que van a trabajar con programas pequeños. No obstante si estos programas fueran obtenidos a partir de proyectos reales y no de libros de programación, como ocurre en la mayoría de los casos de la Tabla 1, los resultados serían más representativos. Por otra parte, los programas utilizados en los estudios revisados contienen pocas faltas, que se han introducido en numerosos caso artificialmente; es decir no son faltas que han ocurrido realmente en el programa sino que se han insertado a posteriori. El utilizar programas de proyectos reales, con faltas reales haría que los resultados de los experimentos se asemejaran más a lo que ocurre en la realidad.

- **Utilizar Variables Respuesta de Interés:** Análogamente a los programas y las faltas, es importante que las variables respuesta del estudio empírico tengan un interés práctico.

  Observando la Tabla 1 se puede concluir que, desafortunadamente, no todos los estudios empíricos se centran en variables respuesta que son de interés para los desarrolladores. Un ejemplo claro de esto es la variable respuesta porcentaje de casos de prueba que detectan al menos un fallo, examinada en [24] [6] [4] [5] [7] . En estos casos los desarrolladores y los investigadores no tienen el mismo concepto de la efectividad de una técnica. Los desarrolladores suelen preferir otro tipo de métricas para la efectividad como el número de defectos detectados por una técnica o el porcentaje de los fallos detectados sobre el total existente.

  Una posible solución a este problema se basa en utilizar el GQM [18] para identificar y definir las métricas a obtener sobre un experimento que sean significativas para la hipótesis que se pretende contrastar y para el contexto del experimento.

- **Aplicación Realista de las Técnicas a Estudiar:** Otro elemento que contribuye a realizar estudios empíricos adecuados es que la propia aplicación de las técnicas a contrastar sea representativa de la realidad. Más específicamente, que la aplicación o uso de las técnicas de pruebas se realice tal como se lleva a cabo en la realidad.

  El papel de las técnicas de pruebas durante el proceso de testing consiste en que los sujetos las aplican (ellos solos o ayudados por algún tipo de herramienta), para obtener un conjunto de casos de prueba que posteriormente es ejecutado y que permite encontrar los defectos del programa. Si observamos la Tabla 1, podemos concluir que solamente tres estudios [1] [11] [25] contemplan aplicaciones reales de las técnicas bajo examen, entendiéndose por esto que en el estudio empírico hay una serie de sujetos que se encargan de aplicar las técnicas en cuestión. El resto de estudios coincide en la generación automática (no intervienen sujetos), principalmente de modo aleatorio, de casos de prueba hasta que se obtiene un conjunto que cumple las restricciones de la técnica. En estos casos, los estudios empíricos están olvidando un tema crucial y es el de los factores humanos. Las técnicas las aplican personas, y puede ocurrir que no todas las personas generen los mismos casos de prueba, e incluso, que no los generen de forma aleatoria, sino siguiendo algún tipo de heurística.

  Sjoberg et al. [21] discuten distintas recomendaciones para proporcionar realismo a los experimentos relacionados con la IS. Entre ellas, destaca el uso de tareas próximas a la realidad en un entorno realista.

- **Proporcionar Paquetes Experimentales que permitan la Replicación:** Una de las características fundamentales que debe tener un estudio empírico es la capacidad de ser repetido exactamente bajo las mismas (o muy parecidas) circunstancias. La existencia de replicaciones permite corroborar o refutar los resultados de un estudio que, de forma aislada, tienen una relevancia muy limitada. Si un experimento no se define con el rigor necesario, dificulta (y en ocasiones hace imposible) su replicación. De hecho, se corre el riesgo de que si la replicación se ha tenido que imaginar y deducir muchos aspectos del experimento, en caso de obtención de resultados distintos a los del experimento original, no se pueda determinar la causa de la diferencia. El problema puede estar tanto en que la replicación no fue tal cual; es decir, se cambió alguna de las condiciones bajo las que se realizó el estudio; o bien puede estar en que, efectivamente, los resultados del estudio original no eran concluyentes y se ha obviado algún detalle en la formulación de hipótesis. La falta de detalle en la descripción del estudio imposibilita determinar bajo qué caso estaríamos.

  La Tabla 1 muestra que la falta de información que conlleva la descripción de los estudios empíricos es un problema bastante extendido. A excepción de [11] [25], los estudios revisados no proporcionan una descripción lo suficientemente exhaustiva como para poder ser replicados, siendo únicamente las variables respuesta las que se explican con suficiente precisión.

  En la comunidad de IS empírica está surgiendo la idea de paquete experimental, como medio para recoger detalladamente todas las condiciones en las que un experimento se ha realizado. No existe todavía, un acuerdo sobre el contenido de estos paquetes, aunque en distintos foros internacionales como la red europea ESERNET (Experimental Software Engineering Network, http://www.esernet.org) o el ISERN (International Software Engineering Research Network, http://www.isern.org) se ha discutido este aspecto. Los resultados de estas discusiones han permitido determinar que la información a proporcionar para describir exhaustivamente un estudio empírico debe abarcar, no solo la descripción del estudio (variables respuesta, factores, ...), sino también el material empleado durante el estudio (ya sea material de formación para los sujetos o material empleado por el experimentador) [17] [9] . Una discusión interesante sobre un ejemplo de paquete experimental puede encontrarse en [19] para la realización de replicaciones sobre experimentos realizados con técnicas de lectura en la Universidad de Maryland (USA) y la de Sao Paulo (Brasil).

- **Encadenar Estudios Empíricos** Es interesante mencionar que los estudios empíricos de la Tabla 1 constituyen, con la excepción del último grupo, estudios aislados, no desarrollados a priori para complementar otros estudios empíricos. Por otra parte, es interesante hacer notar que todos los estudios tienen pendiente confirmación de resultados. Esto ocurre por dos motivos: bien porque hay replicaciones de un experimento que conducen a resultados contradictorios (como [22][1] [3], y [1] [11] [25]); o porque no hay dos experimentos con las mismas hipótesis a contrastar como ocurre en el resto de casos.

  Para el caso concreto de experimentos, la realización de un conjunto de ellos relacionados es lo que se conoce como familia [2], siendo un ejemplo los tres últimos experimentos de la Tabla 1. De esta manera se persigue que los estudios empíricos no ocurran de forma aislada, sino relacionados unos con otros, de tal forma que permitan que el conocimiento empírico crezca y se consolide. La creación de familias de experimentos se basa en la complementación. Se pueden realizar así distintos tipos de experimentos dentro de la misma familia: (1) replicaciones exactas de experimentos de la familia; (2) estudios complementarios, con las mismas hipótesis pero sobre otras técnicas; (3) profundizaciones en las que varían las hipótesis como consecuencia de resultados encontrados en estudios previos [8].

  La idea es, entonces, que cuando un investigador se plantea realizar experimentos analice previamente qué estudios existen sobre el mismo tema o familia y qué tipo de estudio está pendiente para contribuir de manera adecuada a ampliar el cuerpo de conocimientos de la disciplina en cuestión.

- **Secuencia Metodológica de la Experimentación:** Por otra parte, el proceso de consolidación de conocimiento empírico requiere la realización de distintos tipos de estudios, desde experimentos controlados hasta otros estudios realizados en entornos reales. Todo ello debe seguir cierta lógica, por ejemplo, parece razonable realizar varios experimentos controlados antes

de realizar un caso de estudio, ya que los primeros son menos costosos que el último, y parece lógico que se requiera cierta confianza en la veracidad de la hipótesis de partida antes de emplear más recursos.

En la Tabla 1 se puede observar que la secuencia de avance en la madurez del conocimiento empírico en cuanto a los tipos de estudios realizados es bastante caótica. En muchos casos, se realizan casos de estudio antes de haber realizado experimentos o incluso se realizan directamente, sin haber realizado ningún experimento previo.

Los distintos tipos de estudios empíricos se pueden organizar en experimentos y estudios observacionales. Los primeros se caracterizan porque existe, al menos, una variable controlada y pueden ser experimentos controlados o quasi-experimentos según se realicen en un entorno de laboratorio o real [12]. Los estudios observacionales se caracterizan porque no existen variables controladas y pueden ser casos de estudio o estudios de campo. El conocimiento empírico en determinada disciplina se va generando progresivamente, paso a paso a través de estos niveles.

Sería así importante, que los investigadores a la hora de realizar estudios empíricos analizaran los tipos de estudios ya existentes y en qué nivel se encuentran, de manera que los esfuerzos empíricos que se realicen fueran útiles y tuvieran una ubicación significativa en la construcción del cuerpo de conocimientos correspondiente.

## 4. Conclusiones

En este artículo se discuten los problemas de los estudios empíricos que existen actualmente sobre técnicas de testing. Para ello hemos caracterizado los estudios en base a nueve criterios que reflejan la rigurosidad con que se han realizado dichos estudios, su representatividad con respecto a la realidad y su contribución a una estrategia global de contrastación empírica. El análisis de estos criterios muestra importantes carencias que dificultan la creación de un cuerpo de conocimientos sobre técnicas de testing.

Este estudio crítico se ha complementado con la aportación de varias guías que pueden contribuir a paliar los problemas que los estudios empíricos sobre técnicas de pruebas sufren actualmente. Estas guías se refieren fundamentalmente a dos aspectos: (1) la rigurosidad en la realización del estudio empírico, y (2) la necesidad de establecer una serie de acuerdos a nivel de comunidad para coordinar la investigación empírica y así asegurar que los resultados arrojados por los estudios se ratifican y complementan, de tal forma que se pueda avanzar en la creación de un cuerpo de conocimientos sólido y riguroso para tecnicas de testing.

Se ha elegido el área de testing para realizar este trabajo puesto que es uno de los campos de lS en el que existen mayor número de estudios empíricos. Sin embargo, la situación en cuanto a los problemas de los estudios empíricos es similar a en otras áreas, y por tanto las recomendaciones realizadas son igualmente aplicables.

## Referencias


[1] V.R. Basili and R.W. Selby. Comparing the Effectiveness of Software Testing Strategies. Department of Computer Science. University of Maryland. Technical Report TR-1501. College Park. Mayo 1985.

[2] V. Basili, F. Shull and F. Lanubile. Building knowledge through familias of experiments. IEEE Transactions on Software Engineering. Vol 25, N.4. Julio/Agosto 1999.

[3] J.M. Bieman and J.L. Schultz. An Empirical Evaluation (and specification) of the All-du-paths Testing Criterion. *Software Engineering Journal*. Pages 43-51, January 1992.

[4] P. Frankl and O. Iakounenko. Further Empirical Studies of Test Effectiveness. In *Proceedings of the ACM SIGSOFT International Symposium on Foundations on Software Engineering,* Páginas 153-162, Lake Buena Vista, Florida, USA, Noviembre 1998.

[5] P.G. Frankl, S.N. Weiss and C. Hu. All-Uses versus Mutation: An Experimental Comparison of Effectiveness. Polytechnic University, Computer Science Department. Technical Report. PUCS-94-100. Febrero 1994.

[6] P.G. Frankl and S.N. Weiss. An Experimental Comparison of the Effectiveness of the All-uses and All-edges Adequacy Criteria. *Proceedings of the Symposium on Testing, Analysis and Verification*. Páginas 154-164. Victoria, BC, Canada. Octubre 1991.

[7] M. Hutchins, H. Foster, T. Goradia and T. Ostrand. Experiments on the Effectiveness of Dataflow- and Controlflow-Based Test Adequacy Criteria. *Proceedings of the 16th International Conference on Software Engineering*. Páginas 191-200. Sorrento, Italy. IEEE. Mayo 1994.

[8] N. Juristo, A. Moreno. Basics of Software Engineering Experimentation. Kluwer Academmic Publishers USA, 2001.

[9] N. Juristo, A. Moreno, S. Vegas. Could Experiment Packages be a Register for Knowledge Growing?. ISERN 2002. Japon.

[10] N. Juristo, A. Moreno, S. Vegas. Reviewing 25 Years of Testing Technique Experiments. Empirical Software Engineering, vol 9. Páginas 7-44, 2004.

[11] E. Kamsties and C.M. Lott. An Empirical Evaluation of Three Defect-Detection Techniques. *Proceedings of the Fifth European Software Engineering Conference*. Sitges, Spain. Septiembre 1995.

[12] O. Laitenberg, D. Rombach. (Quasi-Expeimental Studies in Industrial Settings. In Lecture Notes on Empirical Software Engineering. World Scientific Vol 12, 2003. Páginas 133-166.

[13] B.F.J. Manly. Multivariate Statistical Methods – A primer. Second Edition. Chapman & Hall, Londres 1994.

[14] A. M. Moreno, S. Vegas. Limitaciones sobre el Conocimiento Empírico de Técnicas de Prueba. VII Jornadas de Ingeniería del Software y Bases de Datos. El Escorial, España, Noviembre 2002

[15] A.J. Offut, A. Lee, G. Rothermel, RH. Untch and Zapf. An Experimental Determination of Sufficient Mutant Operators. *ACM Transactions on Software Engineering and Methodology*. Volume 5 (2). Páginas 99-118. April 1996.

[16] A.J. Offut and S.D. Lee. An Empirical Evaluation of Weak Mutation. *IEEE Transactions on Software Engineering*. Vol. 20(5). Pages 337—344. August 1994.

[17] T. Sandelini, M. Vierimaa. Empirical Studies in ESERNET. In Empirical Methods and Studies in Software Engineering. Experiences from ESERNET. Lecture Notes in Computer Science, 2765, 2003. Páginas 39-54.

[18] R. van Solinger, E. Berghout. The Goal/Question/Metric Method. McGraw Hill, 1999. Japanese Ethnology, Human Organization, Vol 56 (2). Páginas 233-237.

[19] F. Shull et al. Replicating Software Engineering Experiments: Addressing the Tacit Knowledge Problem. 2002 International Symposium on Empirical Software Engineering. 3-4 Octubre, 2002, Nara Japón.

[20] F. Shull, et al. Replicated Studies: Bulding a Body of Knowledge about Software Reading Techniques. In Lecture Notes on Empirical Software Engineering. World Scientific Vol 12, 2003. Páginas 39-84.

[21] D. Sjobert et al. Challenges and Recommendations When Increasing the Realism of Controlled Software Engineering Experiments. In Empirical Methods and Studies in Software Engineering. Experiences from ESERNET. Páginas 24-38. Lecture Notes in Computer Science, 2765, 2003.

[22] E.J. Weyuker. The Cost of Data Flow Testing: An Empirical Study. *IEEE Transactions on Software Engineering*. Volume 16 (2). Pages 121—128. 1990.

[23] C. Wohlin et al. Experimentation in Software Engineering – An Introduction. Kluwer Academic Publishers, USA 1999.

[24] E. Wong and A.P. Mathur. Fault Detection Effectiveness of Mutation and Data-flow Testing. *Software Quality Journal*. Volume 4. Páginas 69—83. 1995.

[25] M. Wood, M. Roper, A. Brooks and J. Miller. Comparing and Combining Software Defect Detection Techniques: A Replicated Empirical Study. *Proceedings of the 6th European Software Engineering Conference*. Zurich, Switzerland. Septiembre 1997.