様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ９ 月 ２２ 日現在

機関番号：８２６２６
研究種目：若手研究(B)
研究期間：2014～2015
課題番号：２６８７０８９９

研究課題名（和文）「測定の不確かさ」の情報がある場合の試験所間比較における統計的方法

研究課題名（英文）Statistical method for proficiency tests with uncertainty information

研究代表者
城野　克広（Shirono, Katsuhiro）

国立研究開発法人産業技術総合研究所・物質計測標準研究部門・主任研究員

研究者番号：６０５０９８００

交付決定額（研究期間全体）：（直接経費）　　1,100,000円

研究成果の概要（和文）：本研究では、試験所間比較のデータを測定の不確かさの情報を用いて解析し、試験所の測定パフォーマンスを評価する方法を提案する。試験所間比較とは同一の品目を異なる試験所が測定することである。測定の不確かさとは測定値の信頼性を定量化したものである。提案手法は２つの段階に分けられる。まず、モデル選択を通じた試験所間比較の企画・運営の妥当性確認が行われる。次に、選択されたモデルに基づく測定パフォーマンスの評価が行われる。高速で十分な精度をもつアルゴリズムの開発によりはじめて実現される、この解析方法は、試験所の技能を公平に評価する一つの手立てを与えるものとなる。

研究成果の概要（英文）：We propose the novel method to evaluate performance of testing laboratories through an interlaboratory comparison, given uncertainty information. An interlaboratory comparison indicates a test in which some laboratories measure an identical item. An uncertainty means the quantified value of the reliability of reported measurement value. In this study, a two-step analysis method is proposed. In the first step, the model selection technique is applied to check the validity of the planning and the implementation of an interlaboratory comparison. In the second step, the performances of the laboratories are evaluated based on the selected model. The fast and precise computation algorithm is developed for the both steps. This analysis could be helpful to secure the fairness in the proficiency evaluation of a testing laboratory.

研究分野：情報学・複合領域

キーワード：技能試験　不確かさ　En数　モデル選択

１．研究開始当初の背景

試験は、多くの市場に出回る商品に対して行われている。自動車の燃費のように多くの人が気にかける試験結果もある一方で、食品表示や電器類の安全試験など通常は目に留まらないところでも多くの試験が行われている。言うまでもなく、多くの試験は試験所で実施される。試験所が正確な試験値を報告することは、健全で公平な試験・検査市場の形成を促し、安心・安全や商取引における公平さを技術的な観点から担保するために重要である。

試験の信頼性を確保するため、近年、「試験所間比較」を介して試験所の技能を確かめ、試験所同士の相互の承認あるいは第三者による認定に役立てる取り組みが進められている。試験所間比較とは同じ品目についての測定を異なる試験所が行い、報告された値を比較・評価することである。

これまで、試験所の技能が十分であることの証明の手立てとして、試験器や標準の校正が行われてきた。これはより高い技能を持つ測定者との１対１の比較であると言える。（左図１Ａ）このように、上位の校正機関との比較の連鎖により、測定結果が国家標準にまで関連づけられていることは、計量計測トレーサビリティと称される。これは従来から、試験結果の信頼性の確保のために求められてきたことである。

近年では、計量計測トレーサビリティに加え、同程度の技能を持つ試験所同士で試験所間比較を実施することも注目されている。（図１Ｂ）試験所間比較を通して、試験所の測定結果が妥当なものか評価することをパフォーマンス評価と呼ぶ。研究目的で行われる試験所間比較と区別するために、パフォーマンス評価の目的で行われる試験所間比較

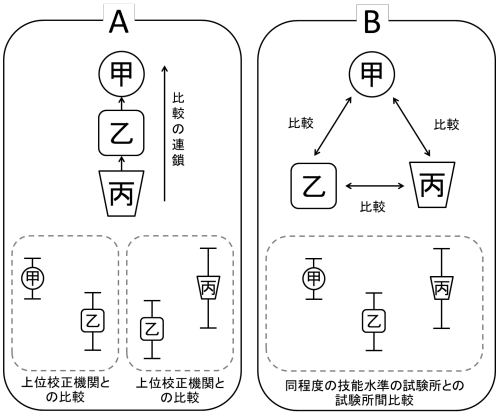

図１　校正の考え方と試験所間比較の考え方の違い。Ａが校正、Ｂが技能試験の考え方にあたる。従来求められてきた校正による品質保証に加えて、試験所間比較による技能の確認が試験所に求められている。

は特に「技能試験」と呼ばれる。例えば、技術的・管理的側面から技能試験を十分な品質で実施するための要求事項は ISO/IEC 17043 **として** 2011 **年に発行された。試験所認定の分野での、技能試験の重要性を示す一つの例であろう。**

これまで試験分野においては、技能試験では測定の不確かさの情報を用いない場合が多かった。測定の不確かさは、測定値の信頼性を定量化したものである。本研究の実施期間中である 2015 **年に発行された** ISO **規格「**Statistical methods for use in proficiency testing by interlaboratory comparison**（試験所間比較による技能試験のための統計的方法、国内ではこの** 2008 **年版のものが** JIS Z 8405 **として発刊されている）」では、参照となる試験所が存在する場合、参照試験所と１対１でパフォーマンス評価する手立ては記載がある。しかし、参加者が同程度の技能の試験所のみで構成される場合のパフォーマンス評価の手立てについては記載がない。**

技能にばらつきのある試験所間比較では外れ値が含まれることが避けられず、ロバストな統計手法を使う必要がある。しかし、従来よく用いられてきたロバストな統計手法では、個々のデータに異なる信頼性の情報、つまり測定の不確かさが付されていることは想定されていなかった。近年、普及の進む測定の不確かさの情報を用いて、試験所の技能を評価するには、これまでにない統計モデルによる解析が必要である。

本研究の実施者らは以前の研究（K. Shirono, Metrologia, vol. 47, 2010, pp. 444-452.**）では、報告された不確かさを拡張することで解析ロバスト性を与えることを実現した。この方法では不確かさは信頼できるパラメータではなく、参考的情報として扱われた。どれほど不確かさを拡張するかは、「ベイズ推定」に基づいて定めた。この中では、どのような統計モデルをベイズ推定に適用するかという点については、技術的知識に従って決定するものとした。この目的のために、現実的な事象に対応するべく、いくつかの統計モデルを提案し、その解析結果を比較することをしてきた。実施者以外の従来研究も同様のアプローチのものが多い。**

しかし、実際のところ、統計モデルを技能試験の供給者が決定することに困難さがあることは否めない。統計関連の研究分野において応用の進む「モデル選択」を、本研究のテーマに当てはめ、任意性の少ない解析手法を提案することは、自然な研究の拡張であった。しかし、その実施のためには、パラメータ数が多いこと（最大で数百）と、確率密度関数が複雑な形状になりやすいことから、計算面での困難があった。実施者は応募時点で、一部の条件では積分計算を近似により解析的に実施できることを確かめた。この点では、一定の見通しの下で研究は開始された。

２．研究の目的

本研究の目的は、測定の不確かさの情報を用いて、試験所間比較により試験所のパフォーマンスを評価するための方法を提案することにある。実用可能な試験所間比較による試験所のパフォーマンスの評価方法を与えるべく、以下のことを明らかにすることを目的とする。

１．　試験所間比較の企画・運営がパフォーマンス評価に適したものであることチェック

２．　個々の試験所のパフォーマンスの統計モデルに基づく評価

１．について説明する。試験所間比較の結果、図２Ａのように不確かさの範囲での一致が全く見られない場合には、回付された品目の安定性が低く、回付の途中で特性が変化した可能性もある。もしその統計モデルが妥当ならば、試験所間比較の企画・運営に試験所のパフォーマンスを評価するだけの品質がないことを意味する。モデルの選択を通じ、試験所間比較のプロトコルが妥当であるかを統計学的見地から検証する。

２．について説明する。試験所間比較の結果、図２Ｂのようになった場合に、丙の技能を評価するにあたっては、「丙の値は甲、乙、丁のうち、十分な技能をもつものが報告した値と同じ平均値を持ち、丙の報告した測定の不確かさをばらつきとする分布に従う」という仮説を検定する。この仮説を検定するために、どのような統計量を用いればよいかを決定する。

この２つの研究目的は、本研究を特色づける。実施者らが報告した以外にも多くの従来研究が報告されているが、今回提案する研究内容は、モデル選択を通じた試験所間比較の企画・運営の妥当性確認と、それと一貫したモデルに基づく技能の評価を実施する点に独自性がある。実施者の知る限りでは、従来の研究は「試験所間比較の企画・運営が妥当でないときにどうするか」という視点のものと、「技能の低い試験所をどう見出すか」という視点のものしかなかった。高速で十分な精度をもつアルゴリズムの開発によりはじめて実現される本研究の特色は、パフォーマンスの評価の客観性を高める上で、強力な利点となる。

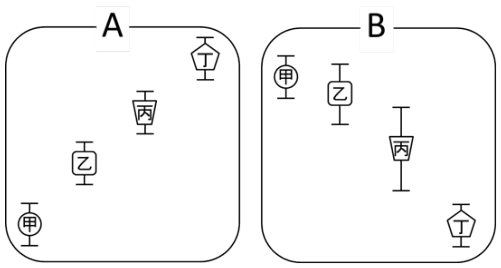

図２　本研究の目的は大きく２つある。報告された結果が不確かさの範囲を大きく超えてばらつくとき、技能試験の企画・運営に問題あるかも知れない。これをチェックする（A）。丙の値が甲、乙と不確かさの範囲で一致しているかどうかを、外れた値である丁に影響されずに決める（B）。

さらに言うと、現実的な時間（例えば１０分程度）で解析が終了することが実用上は求められる。特に、モンテカルロ計算を積分に使用すると、数日以上かかることが予想される。十分な精度の高速に計算できる近似アルゴリズムを開発することも目的のひとつと数えられる。

３．研究の方法

（１）　試験所間比較の企画・運営がパフォーマンス評価に適したものであることチェック

この目的のために、大きく分けて２つの統計モデルを用いる。一つはどんなに全体の中心に近い報告値にも追加的不確かさを含む統計モデルであり、もう一つは中心から一定以上離れた値にのみ外れ具合に応じた追加的な不確かさを与える統計モデルである。一致が比較的よい試験所でさえ追加的不確かさの存在が考えられるということであれば、試験所間比較の企画・運用が十分に妥当なものでないことを意味している。具体的には、測定対象の安定性が低く事実上同一の量が測定されていないことや、同一の性質を持つと考えて配付された測定対象の均質性が低く、その不均質性が重要な効果を有していることが考えられる。

試験所 i が報告値 x_i とその不確かさとして標準偏差にあたる値（標準不確かさと呼ぶ）u_i を報告したものとする。上のことを実現するために、具体的には、$x_i \sim \mathrm{N}(\mu, u_i^2+\sigma_c^2)$という統計モデル（モデルＣと呼ぶ）か、$x_i \sim \mathrm{N}(\mu, u_i^2+\sigma_i^2)$という統計モデル（モデルＩと呼ぶ）のどちらかを、すべての試験所に当てはめる。ここで、$\mathrm{N}(\mu, \sigma^2)$は平均$\mu$で分散$\sigma^2$の正規分布を表す。$\sigma_c^2$が上に述べた一致が比較的よい試験所に与える追加的不確かさによる分散である。σ_i^2は値の一致が悪い試験所が解析に影響を与えないように、付加する追加的不確かさの分散である。σ_c^2はモデルＣを与えたすべての報告値に共通である。σ_i^2はモデルＩを与えたすべての報告値に異なる値が与えられる。この統計モデルにより、ベイズ推定に必要な尤度が与えられる。

事前分布を適切に選ぶことで、モデル選択とベイズ推定が可能になる。モデル選択はそのモデルの周辺尤度の比較によって実施する。周辺尤度とは、簡単に言えば、統計モデルの尤度のことである。周辺尤度が大きい統計モデルほど、尤もらしいモデルと言える。もし、すべての報告値について、モデルＩが選ばれれば、最も一致性の高い試験所については、σ_i^2が十分に小さく評価される。これは、

少なくとも良いパフォーマンスの試験所の報告値に、付加的な不確かさを考慮する必要がないことを意味する。つまり、報告した値に修正をせずとも、一部のデータについては、十分にそのばらつきを説明できるということである。一方で、もし、２つ以上の報告値にモデルＣを選ぶことがあれば、その試験結果を説明するには統計モデルの修正が必要であるということになる。このことから、試験所間比較の企画・運用に問題あることが示唆される。（なお、１つの報告値にのみモデルＣが選ばれることは、すべての報告値にモデルＩを選ぶことと等価であるので考えない。）

本研究で、我々は周辺尤度を最大にするように、事前分布を選ぶこととした。しかし、事前分布として、柔軟過ぎる関数形を用いると我々の目的とする解析が不可能となる。任意の関数形から、周辺尤度を最大にするように事前分布を選ぶと、その関数は必ずデルタ関数となる。モデルＩに対して、デルタ関数を事前分布の関数形として適用した場合について、我々はまず報告した［主な発表論文等〔図書〕①］。この報告の中では、図２Ａに示したようなデータに適用した際に、不合理なパフォーマンス評価結果が得られることを報告した。加えて、デルタ関数を事前分布として、モデルＣについても検討すると、いずれかの報告値についてモデルＣを選んだとき、すべての報告値にモデルＩを選んだときより大きい周辺尤度を与えることがないことが、数学的検討により明らかとなる。すなわち、デルタ関数は我々の目的を達成するための事前分布になり得ない。少なくとも、不確かさの範囲で値が全く一致しないデータでは、モデルＣが選ばれるような事前分布の関数形を慎重に選ぶ必要がある。

そこで、我々はべき乗の事前分布を適用することを提案した。具体的な関数は、［主な発表論文等〔雑誌〕①］に示す。この関数形を導入することにより、

１．　極めて値の一致が悪いデータにはすべての試験所でモデルＣが選ばれる。
２．　極めて値の一致が良いデータに対してはすべての試験所でモデルＩが選ばれる。
３．　ほとんどの場合で、積分計算を近似的なアルゴリズムで高速に実現できる。

という特徴を導くことができる。

もちろん、この手法が実用的であるためには、解析がロバストに実施される必要がある。ロバストな解析が実現されているかは、最終的に選ばれた統計モデルにより推定されるμの値が、その試験所間比較を代表する値として不自然でないことにより確かめられる。実のところ、４章に示すように、ロバストな解析が可能である。後の便宜のために、ここで得られたμの値を「ロバストに得られた平均」と呼ぶ。

（２）　個々の試験所のパフォーマンスの統計モデルに基づく評価

技能試験の企画・運営がパフォーマンス評価に適したものであることチェックと、個々の試験所のパフォーマンスの統計モデルに基づく評価が、少なくとも同じような統計モデルに基づいていることは、解析の一貫性のために重要なことである。

まず、計算を簡単に実施するために、技能試験の企画・運営のチェックにおいて得られた統計モデルにおいて、未知パラメータの大部分を推定値に置き換えることを考える。さらにここでは、ベイズ統計を用いずに最尤推定によるパラメータ推定を与えることを考える。ただし、統計モデル $x_i \sim \mathrm{N}(\mu, u_i^2+\sigma_i^2)$を用いた単純な最尤推定はロバストに得られた平均と著しく異なるμの推定値が与えられ、一貫性を欠くことがある。そこで、我々は尤度のローカルマキシマムの中から、最もロバストに得られた平均に近い値を示す尤度のピークを見出し、そのピークに属するσ_i^2 を$\sigma_{i\text{-LML}}^2$として、今回の目的に使用する。

この$\sigma_{i\text{-LML}}^2$を用いて、試験所 k のパフォーマンスについて考えてみる。試験所 k のパフォーマンスが満足なものであるということは、$x_k \sim \mathrm{N}(\mu, u_k^2)$という統計モデルの妥当性が確認できるということと同義であろう。$\sigma_{k\text{-LML}}^2$がゼロになっているかどうかは興味のあるところであるが、仮に$\sigma_{k\text{-LML}}^2$がゼロでない時、それを根拠に、$x_k \sim \mathrm{N}(\mu, u_k^2)$という統計モデルを棄却できるかは疑問が残る。そこで、我々は $x_k \sim \mathrm{N}(\mu, u_k^2)$、$x_i \sim \mathrm{N}(\mu, u_i^2+ \sigma_{i\text{-LML}}^2)$ $(i \neq k)$という統計モデルを考えた。この統計モデルは技能試験の企画・運営のチェックにおいて得られた統計モデルに近いと言える。もし、このモデルが棄却されれば、$x_k \sim \mathrm{N}(\mu, u_k^2)$という統計モデルが妥当でないとういことを意味することになる。

この検定のために、我々は ISO/IEC 17043 に記載にある“参照となる試験所が存在する場合、その試験所と１対１でパフォーマンス評価する手立てである“ E_n 数 ”との類似性を考えて、新しい統計量を導いた。具体的な形式は、［主な発表論文等〔雑誌〕①］に示す。論文の中では、他のベイズ推定を用いた点推定量を使った場合も試し、同程度の性能を示すことを明らかにした。

このように、確かめたい量以外の未知パラメータを最尤法やベイズ統計から得られる値に固定することで、数値的な積分計算はもはや不必要となる。100 程度の試験所が参加する場合、未知パラメータの決定のステップには、3 GHz の PC で 10 分～20 分程度の時間がかかるが、パフォーマンス評価のステップでの計算時間は事実上無視できるほどになる。この計算プログラムは、［主な発表論文等〔雑誌〕①］および［主な発表論文等〔雑誌〕②］に電子媒体で付録として提供した。

４．研究成果

今回開発した方法を、参照試験所が実際には存在する技能試験データについて、模擬的に適用した。その中では、興味深い結果を得ることができた。ここでは、議論の簡便さのために、シミュレーションデータに提案手法を適用し、その結果を報告する。

本研究の手法を図３のデータに適用する。詳細な数値は表１に示す。このデータにおいては、x_1 および x_6、x_7 が極端な外れ値となっている。$x_3 \sim x_5$ は同一の値である。x_2 は、$x_3 \sim x_5$ の値よりやや小さいが、報告した標準偏差の２倍の範囲にはその値を含む。このデータの様相からは、試験所 2 ~ 5 のみパフォーマンスが良いものと見受けられる。

まず、技能試験の企画・運営がパフォーマンス評価に適したものであることのチェックを行う。すべての報告値にモデルＩを適用した場合の周辺尤度は、すべての報告値にモデルＣを適用した場合の周辺尤度の 10.6 倍大きな値になった。また、このケースでは一部の報告値にのみモデルＣを適用した場合は、すべての報告値にモデルＣを適用した場合よりも周辺尤度は小さかった。このため、すべての報告値にモデルＩを適用するのが、統計学的見地からは最も可能性が高い統計モデルであると言える。すなわち、技能試験の企画・運営がパフォーマンス評価に適しており、技能試験品目の不安定性などの問題は小さいと言える。また、平均μは3.97と推定される。その推定の標準偏差は0.17である。この値は $x_3 \sim x_5$ とよく一致すると言ってよく、ロバストな解析が実現されていることを示している。

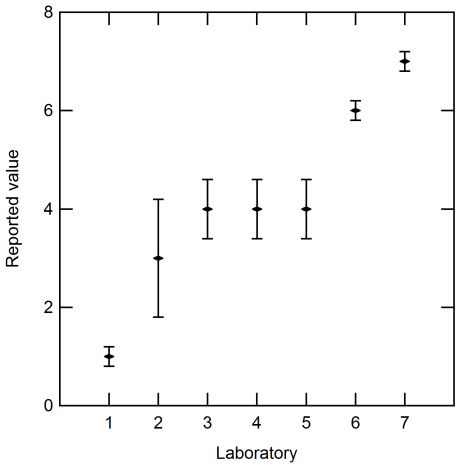

図３　本研究での提案手法の性能を確かめるための模擬データを示した図。図中の縦線は報告された標準偏差の２倍の大きさを意味する。

表１　計算例のデータと解析結果。ここでの E_n 数は本研究で提案した従来の E_n 数を拡張したもの。その絶対値が１より大きい場合、パフォーマンスは不満足とされる。

	x_i	u_i	$\sigma_{i\text{-LML}}{}^2$	E_n数
試験所1	1	0.1	8.9	−7.6
試験所2	3	0.6	0.6	−0.8
試験所3	4	0.3	0	0
試験所4	4	0.3	0	0
試験所5	4	0.3	0	0
試験所6	6	0.1	4.1	5.1
試験所7	7	0.1	9.1	7.7

本研究の手法に基づくローカルな最尤法により推定したパラメータ $\sigma_{i\text{-LML}}{}^2$ を表 1 に示す。i = 3, 4, 5 には $\sigma_{i\text{-LML}}{}^2 = 0$ となる。これらの報告値については付加的な分散を付与しないことが最も尤もらしいことを意味している。他の試験所の値については、付加的な分散の最尤値は正の値という結果となる。

表１には本研究で提案した手法で計算した E_n 数も示す。E_n 数は、その絶対値が１より小さい場合と大きい場合で、パフォーマンスはそれぞれ「満足」、「不満足」とされる。試験所 1 の E_n 数は−7.6 と, その絶対値は 1 より大きくなっており、「不満足」なパフォーマンス評価が与えられる。試験所 6、7 も同様に 1 より大きな E_n 数となり、パフォーマンスは「不満足」である。

試験所 3 ~ 5 に着目すると、E_n 数は 0.0 となり、「満足」なパフォーマンス評価が与えられる。このように、$\sigma_{i\text{-LML}}{}^2 = 0$ となる場合には、パフォーマンス評価は「満足」と与えられる。これは受け入れやすい結果であろう。試験所 2 の結果に着目すると、$\sigma_{i\text{-LML}}{}^2 = 0.61 > 0$ と与えられるにも関わらず、E_n 数は−0.8 となり、「満足」なパフォーマンスが与えられる。このように、$\sigma_{i\text{-LML}}{}^2 > 0$ なるときに、パフォーマンス評価が「不満足」となるとは限らない。最尤法によるパラメータ推定には、不確かさが伴う。本研究の提案手法によるパフォーマンス評価は、その不確かさも加味したものと解釈できる。

このように実際の例に近い例において、妥当に解析が行われることを確かめることができた。このデータの他、多くの技能試験データに適用したが、ほとんどのケースで開発したアルゴリズムは安定的に動作し、その与える結果は妥当なものであった。さらに言えば、最尤法の適用可能性を明らかにしたことは、極めて簡便な評価手法が与えられる道筋を示したものと言える。

製品安全のルールを国際的に共通化することが、昨今社会的に大きな話題となっている。その基盤として、測定の信頼性を確保することが求められる。食品、環境、臨床検査の分野でも、同じ状況にある。これらの分野の試験・検査の国内市場規模はそれぞれに年数千億〜数兆円とされる。本研究の手法が現

実に適用されることにより、適切な評価を通じ、高い技能の試験所に市場での優位性を与えることになるだろう。それは、安心・安全や商取引における公平さを技術的な観点から担保することにつながる。多くの分野で、高い試験品質を有する日本において、この研究課題で、世界をリードする意義は大きい。

５．主な発表論文等

〔雑誌論文〕（計　２　件）

① K. Shirono, M. Shiro, H. Tanaka, K. Ehara, Proficiency tests with uncertainty information: Extension of the E_n number for cases with no reference laboratory 、Measurement、査読有、vol. 83、2016、pp. 135-143.
DOI：10.1016/j.measurement.2015.09.035

② K. Shirono, M. Shiro, H. Tanaka, K. Ehara, Proficiency tests with uncertainty information: Detection of an unknown random effect、Measurement、査読有、vol. 83、2016、 2016, 144-152.
DOI: 10.1016/j.measurement.2016.01.002

〔学会発表〕（計　５　件）

① 城野克広、城真範、田中秀幸、榎原研正、ベイズ統計に基づく不確かさのある試験所間比較による技能試験の解析方法、日本品質管理学会第 104 回研究発表会（東京）

② K. Shirono, M. Shiro, H. Tanaka, K. Ehara, Theory and computation tool for the novel KCRV and DOE determination method based on Bayesian statistics, Advanced Mathematical and Computation Tools for Measurement and tensting X（サンクトペテルブルグ、ロシア）

③ 城野克広、城真範、田中秀幸、榎原研正、ベイズ統計に基づく不確かさのある技能試験におけるパフォーマンス評価、日本品質管理学会第 107 回研究発表会（東京）

④ K. Shirono, M. Shiro, H. Tanaka, K. Ehara, Proficiency test with the information of uncertainty: analysis with the maximum likelihood, IMEKO XXI World Congress（プラハ、チェコ共和国）

⑤ 城野克広、海野泰裕、米沢仲四郎、ガンマ線スペクトロメトリーにおける検出効率曲線の決定とその不確かさ評価、日本分析化学会第 64 年会（福岡）

〔図書〕（計　１　件）

① K. Shirono, M. Shiro, H. Tanaka, K. Ehara, Advanced Mathematical and Computational Tools in Metrology and Testing X、査読有、pp. 357-368.

〔その他〕
ホームページ等
https://staff.aist.go.jp/k.shirono/

６．研究組織

(1)研究代表者
城野克広（SHIRONO, Katsuhiro）
産業技術総合研究所・物質計測標準研究部門・主任研究員

研究者番号：60509800