総説

学習者評価の信頼性とは

大西　弘高
東京大学医学系研究科医学教育国際研究センター

抄録

評価手法が良いか否かを議論するためには妥当性が重要だが，総括評価に客観試験以外の評価手法を用いる際，信頼性検証が不可欠である．OSCE（objective structured clinical examination）における信頼性指標は，各課題に挙げられている項目毎の評価者間信頼性と，ステーション毎にまとめた項目合計点をステーション間でみたときの信頼性に大別される．評価者間の信頼性指標には，評価者間一致率，κ 値，級内相関係数が挙げられる．ステーション間の信頼性は α 係数で表すことが可能だが，一般化可能性理論を用いることで，誤差要因同士の分散の比較，一般化可能性係数や信頼度指数の計算，決定解析が可能となる．

キーワード：学習者評価，信頼性，テスト理論，一般化可能性理論

Reliability in learner assessment

Hirotaka Onishi
International Research Center for Medical Education, Graduate School of Medicine, The University of Tokyo

Abstract

Validity is important to discuss whether the assessment tool is good or not, evaluation of the validity is important. However, identifying reliability is indispensable when non-objective assessment tools are used for summative purposes. Two types of reliability indices are used for OSCE（objective structured clinical examination）; one is the inter-rater reliability for each item of the checklist in a station task, and the other is inter-station reliability using the total item scores from all the stations. Indices of inter-rater reliability consist of inter-rater agreement, kappa value, and intra-class correlation coefficients. Inter-station reliability can be expressed by the alpha coefficient. Utilization of the generalizability theory in OSCE enables comparison of variances among different error factors, calculation of the generalizability and dependability indices, and decision study.

Key words：learner assessment, reliability, test theory, generalizability theory

はじめに

学習者を評価する際，精神運動領域，情意領域，高次認知領域の評価等においては評価者のバラツキの懸念が大きくなる．総括評価，特に high-stake test においては，一定以上の信頼性が達成されていなければ，合否判定等に関する誤差が大きくなるため，信頼性を検証し，改善することは不可欠といえるだろう．

本稿では，OSCE による評価をイメージしながら信頼性指標の持つ意味，様々な信頼性指標の意義に分け

【連絡先】東京大学医学系研究科医学教育国際研究センター
〒113-0033 東京都文京区本郷 7－3－1 医学部総合中央館 2 階
受理日：2018年 2 月 9 日　採録決定日：2018年 3 月 6 日

表1　妥当性を証明するための情報源

内容	テストで得たスコアの処理プロセス	テストの内的構造	他の変数との関係	結果
・試験のブループリント，達成すべき目標領域，テスト内容が一貫している ・テストが目標領域を代表している ・テストで問われている内容が現場で必要となる内容と一致している ・テストの問題の質が高い ・問題作成者が一定レベルの質に達している	・学習者がテストのフォーマットに慣れているか ・採点が正確かどうかの確認 ・異なったフォーマットの点数を正確に合計しているか ・スコアや評価の正確性に関する質管理 ・スコアと合否判定に関する質管理 ・学習者や教育者への報告に関する質管理 ・スコアの記述が分かりやすく，正確か	・項目分析データ（項目の難易度，識別係数，項目特性曲線，項目間相関，項目と合計点との相関） ・項目の信頼性，測定の標準誤差，一般化可能性分析，因子分析 ・心理測定モデル	・他の関連した変数との相関 ・類似したテストとの収束的相関（内的，外的） ・異なったテストとの弁別的相関（内的，外的） ・テストと規準（クライテリア）との相関 ・エビデンスの一般化可能性（外挿可能性）	・テスト結果が学習者や社会に与える影響 ・将来の学習者に与える影響 ・正の結果が予測しなかった負の結果を上回っているか ・合否基準となるスコアを決定する方法が納得できるものか ・合否決定による分類の正確性 ・能力が不十分なのに合格になったり，能力が十分なのに不合格になったりする者の数 ・合否判定基準スコアの条件付き標準誤差 ・評価が学習者に与える影響

て概説し，例を挙げてその有用性にも触れてみたい．

信頼性指標の意味

評価には常に信頼性，妥当性の問題が付いて回る．妥当性とは，その評価手法が真に測定したいものを測定できているかどうか，信頼性とは，その評価手法の測定が安定しているかと定義される．学習者評価が学習者の能力を反映しているかどうかは，妥当性で判断されるべきであり，信頼性の問題はさほど大きくないはずである．

妥当性は，Messick の再定義により，構成概念妥当性に一元化されると共に内容がやや拡がり，テストデータの処理，結果が社会に与えるインパクト等も考慮することになった[1]．Downing はこの新しい定義に基づき，妥当性を証明するための情報源を**表1**のように列挙した[2]．特に，合否判定に関する内容は以前よりもはるかに重視されるに至った．

さて，再度各学生の評点の持つ意味を考えてみよう．学生の評点が60点であった場合，これはいくらかの誤差を含んだ点推定値であるとみなすことも可能である．ここで，測定値の標準誤差（standard error of measurement）が5点で，評点が正規分布に従うと仮定すれば，［50.2（60－5 ×1.96）～69.8（60+ 5 ×1.96）］が点推定値の95% 信頼区間となる．体重計に乗って60kg を指したとき，その95% 信頼区間が50～70kg であったなら，その体重計は全く信頼できないと呆れるだろうが，学習者評価の評点というのはそのような性質の数値であると認識しておくべきだろう．

ところが，合否を判定しようとすれば，60点の学生が合格で59点の学生が不合格，ということも起こりうる．このとき，点推定値としての評点の信頼区間があまりに広いと，特にぎりぎりで不合格になった受験者から試験に疑義を挟まれる可能性はある．そのような問題に可能な限り対処するために，学習者評価の信頼性に配慮が必要なのである．

まずは，信頼性指標とは何かを明確にしていこう．概念的には，同じ人が違うタイミング，違う視点で何回も評点を付ける，あるいは異なる人が評点を付けるといった場合に，その評点がどの程度安定しているかを意味する．

$$\text{信頼性係数} = \frac{\text{真値の分散}}{\text{真値の分散} + \text{誤差分散}} \quad \cdots \text{式①}$$

式①より，誤差分散が小さくなると信頼性係数が大きくなる関係があることが分かる．各施設レベルの試験なら0.7，全国レベルの試験なら0.8という信頼性係数が目安として示されている[3]．

表２　信頼性指標の種類

１．評価者間信頼性指標
（ア）項目毎の信頼性指標
①　評価者間一致率
②　κ 値
③　級内相関係数（通常は ICC（２，１））
（イ）ステーションの項目合計点を用いた信頼性指標
①　Pearson の相関係数
②　Spearman の相関係数
２．ステーション間の信頼性指標
（ア）Cronbach の α 係数
３．評価者とステーションの要因を併せ持つ信頼性指標
（ア）一般化可能性係数（generalizability index，G係数）
（イ）信頼度指数（dependability index，Φ係数）

表３　２名の評価者によるチェックリスト評点の例

		評価者２		
		0	1	計
評価者	0	0	0	0
1	1	1	49	50
	計	1	49	50

注１：観察された一致率＝ Po，一致率の期待値＝ Pe とすると，Po ＝ 49/50 ＝ 0.98，Pe ＝ 49/50 ＝ 0.98 であり，κ 値＝（Po － Pe）/（１－ Pe）＝０

注２：被評価者数 n，評価者数 k と置き，BMS（between-subject mean squares），JMS（between measures（judges）mean squares），EMS（error mean squares）を用いて表すと，ICC（２，１）＝（BMS － EMS）/{BMS ＋（k － 1）EMS ＋ k（JMS － EMS）/n} となる．文献[4]の定義によって計算すると，BMS ＝ 0.01，EMS ＝ 0.01 であり，ICC（２，１）＝０となる．

信頼性指標の種類やその意義

様々な信頼性指標を**表２**に一括して挙げた．このうち，評価者間の信頼性指標については，項目毎のものとステーションの項目合計点を用いたものに分けられる．ステーション間の信頼性指標を求める際には各ステーションの項目合計点が用いられる[3]．

このうち，利用の意義が大きいのは項目毎の評価者間信頼性指標と信頼度指数（dependability index）である．これらについて，下に詳述する．

項目毎の評価者間信頼性指標

項目毎の評価者間信頼性指標は，各項目がどのような性質を持ち，評価基準が明確になっているかどうかを確認するのに役立つことがある．例えば，医療面接ステーションで「患者の姓を確認した」という項目は評価者間のズレが生じにくく，「患者の話を共感的に聴いた」という項目はズレが生じやすいという違いが明確になるかもしれない．後者に対し「患者の話を聴きながら５回以上相づちを打った」という評価基準を設ければ評価者間信頼性は向上するかもしれないが，マニュアル的な対応を増やして妥当性を減じるかもしれない．

κ 値や級内相関係数は項目の得点率が50% 付近では信頼区間が小さいが，０% や100% に近づくと信頼区間が大きくなるため注意が必要である．例えば**表３**では「患者に挨拶した」という項目に対し，ほぼ全員が正しく実施できているが，ただ１人の学生に対する評価がずれている．このとき，評価者間一致率は0.98だが，κ 値，級内相関係数（ここでは評価者間一致度を表すときに頻用される ICC（２，１）[4] を用いる）のいずれもが０となる．もし，この評価がずれた学生の真の評価は０点であって，評価者１がそれを正しく０点としていたとすれば，全ての指標はいずれも１となる．これらの観察から，評価者間一致率と κ 値，あるいは級内相関係数の併記が奨められている[5]．

なお，評価者が３名以上のときに評価者間信頼性指

表４　OSCE 結果による一般化可能性解析の例

要因	偏差平方和	自由度	平均平方	分散推定値	分散の比率
学生（i）	4.4813	91	0.04924	0.004422	49.5%
ステーション（j）	0.6672	4	0.16680	0.000804	9.0%
評価者（k：j）	0.0801	5	0.01602	0.000151	1.7%
学生×ステーション（ij）	1.8273	364	0.00502	0.001463	16.4%
学生×評価者（i（k：j））	0.9527	455	0.00209	0.002094	23.4%
Sum	8.0086	919		0.008934	

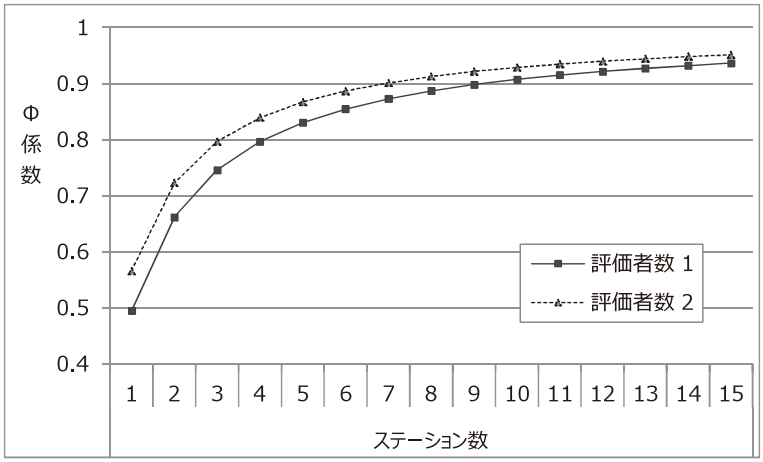

図１　表４と同じデータを用いた決定解析の例

標を求めようとすれば，級内相関係数（intraclass correlation coefficient：ICC）が最も簡便に求められる．ICC にはどの分散を用いるかによって６種類のデータが得られるが，複数評価者を変量モデルとした場合には ICC（２，１）が用いられる[4]．

κ 値は，どの程度の数値なら評価者間信頼性が保たれているかに関するガイドラインがあることでも知られる．Landis and Koch は，$\kappa \leqq 0.2$ なら Poor，$0.2 < \kappa \leqq 0.4$ なら Fair，$0.4 < \kappa \leqq 0.6$ なら Moderate，$0.6 < \kappa \leqq 0.8$ なら Good，$0.8 < \kappa \leqq 1.0$ なら Very good とした[6]．Fleiss は，$0.75 \leqq \kappa$ なら Excellent，$0.40 < \kappa < 0.75$ なら Fair to good，$\kappa \leqq 0.40$ なら poor とした[7]．このガイドラインは ICC にも用いられることがある．しかし，これらの基準には数値的な根拠は特になく，一部の統計学者はこのガイドラインの使用を憂慮している[8]．

評価者間信頼性指標のうち，ステーションの項目合計点を用いたものはさほど意味を持たない．相関係数が低くても，どこを改善すべきかには項目毎の指標が必要になるからである．ステーション間の信頼性指標を求めれば，ステーション数を増やしたときにどの程度信頼性指標が上がるかが分かるため，参考になる．このような分析は α 係数，信頼度指数のいずれでも可能だが，評価の主観性が問題になりがちな OSCE のような場合には，これも含めて解析できる後者がより有用であろう．

一般化可能性解析の概要

一般化可能性解析では各ステーションの項目合計点を用い，分散分析の手法によって式①における誤差分散を３つの要因（学生の評点の差によるもの（i），ステーション（j），評価者（κ））で分離することが主眼となる．その際，各要因の交互作用も得られるが，評価者はステーション毎に異なるため，評価者とステーションの交互作用を求めることはできない（要因のネスト構造と呼ばれ，評価者の分散は $\sigma_{(\kappa:j)}$ と表記される）．よって，学生，ステーション，評価者の分散と，学生とステーションの交互作用の分散，学生と評価者の交互作用の分散の５つが分離されることになる．

信頼度指数Φは，学生の分散と，それ以外の分散成分から，式②の形

$$\Phi = \frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2 + \frac{\hat{\sigma}_{ij}^2 + \hat{\sigma}_j^2}{n} + \frac{\hat{\sigma}_{i(\kappa:j)}^2 + \hat{\sigma}_{(\kappa:j)}^2}{nr}} \quad \cdots \text{式②}$$

（n, r はステーション数及び評価者数）で求まる．なお，

式②において学生要因である i を含まない項を削除したものが一般化可能性係数（generalizability index）となる．一般化可能性係数は相対評価，信頼度指数は絶対評価に利用すべきとされる．詳細については，文献を参照していただきたい[9-11]．

OSCE 評点を用いた一般化可能性解析の例

表4には，受験者92名，5ステーション，各ステーション評価者2名の形で実施されたOSCEの評点を用いた一般化可能性解析（generalizability study）の例を挙げた．ここには，5種類の分散推定値やその割合が示されている．絶対評価に用いることを想定して式②に代入すると，$\Phi = 0.004422/(0.004422 + 0.000804/5 + 0.001463/5 + 0.000151/10 + 0.002094/10) = 0.867$となる．**図1**には，これと同じデータを用いた決定解析（decision study）のグラフを示した．決定解析はステーション数や評価者数を変更したときに，信頼度指数がどう変化するかを予測するものである．例えば，ステーション数7，評価者数2とすれば，$\Phi = 0.004422/(0.004422 + 0.000804/7 + 0.001463/7 + 0.000151/14 + 0.002094/14) = 0.901$となる．

まとめ

評価の信頼性に関する一般的な内容を，OSCEの例を中心にまとめた．特に，項目毎の評価者間信頼性指標，複数の相関要因を含めた一般化可能性解析については，OSCEの信頼性向上に対して有用であり，詳述した．

文献

1）Messick, S.（1989）. Validity. In : Educational Measurement, 3rd edn. Ed : Linn RL. New York : American Council on Education and Macmillan, 13–104.
2）Downing, SM.（2003）. Validity : on the meaningful interpretation of assessment data. Med Educ, 37 : 830–837.
3）Downing, SM.（2004）. Reliability : on the reproducibility of assessment data. Med Educ, 38 : 1006–1012.
4）Hasnain, M, Onishi, H, Elstein, A.S.（2004）. Inter-rater agreement in judging errors in diagnostic reasoning. Med Educ, 38 : 609–616.
5）Shrout, P.E, Fleiss, J.L.（1979）. Intraclass Correlations : Uses in Assessing Rater Reliability. Psychol Bull, 86 : 420–428.
6）Landis, J.R, Koch, G.（1977）. The measurement of observer agreement for categorical data. Biometrics, 33 : 159–174.
7）Fleiss, J.L.（1981）. Statistical Methods for Rates and Proportions, 2nd edn. New York : Wiley, : 212–236.
8）Uebersax, J.（2002）. Kappa Coefficients. http : //ourworld.compuserve.com/homepages/jsuebersax/kappa.htm（accessed 11 October 2006）.
9）Shavelson, R.（1991）. Webb, N. Generalizability Theory ; A Primer. Sage Publications, Thousand Oaks.
10）池田央.（1994）. 現代テスト理論，東京：朝倉書店.
11）Brennan, R.L.（2005）. Generalizability theory, New York : Springer.