大学入試研究ジャーナル第 30 号，248-253，2020.

課題探究の取り組みを多面的に評価する方法

――評価の枠組みと方法の検討――

雨森 聡，宇佐美 壽英，藤井 朋之（静岡大学）

筆者らは 2016 年度から課題探究とその評価に取り組んでいる。本稿は，その取り組みにおいて現れる高校生の能力を評価する方法について説明，提案することを目的としている。評価では，チェックリストとルーブリックを用いる場面があり，これらの妥当性と信頼性について検討している。能力を多面的に評価する方法は，短時間で完璧なものにすることは困難であり，当事者の検討ならびに議論，事前の共通理解の浸透，検討・議論を元にした改善が必要であることがわかった。

キーワード：課題探究，主体性評価，評価の妥当性・信頼性，入試のデザイン

1 本稿の目的とねらい

筆者らは，2016 年度から高大接続的な内容のパフォーマンス課題とその評価を，地域の高校と大学の関係者が協働する高大連携の体制で実践を行っており，その実践の枠組み等については一度まとめている（雨森ほか，2019）。そこでは，具体的な評価方法についてはあえて言及しなかったが，その理由は評価を軽視しているわけではなく，議論の複雑化や焦点がぼけることを回避するためである。

本稿の目的は，筆者らが行っている課題探究の取り組みでの評価方法について説明するとともに，パフォーマンス課題に対する多面的評価の方法について提案することである。先に断っておくが，評価に関する新機軸を打ち出すことを目指してはいない。また，そのような大それたことは筆者らの能力を超える。本稿が，筆者らと同様な取り組みを行っている方々や，行おうとしている方々に対して，とくに評価について困っている方々に対して，ひとつの実践例として参考になれば幸いであると考えている。

なお，本稿で扱う実践は，大学入試で用いることを想定している。高大接続答申で述べられた大学入試の問題点は，知識の評価に偏重し過ぎない，知識の評価を軽視しないことであり，換言すると，学力を多面的に評価することである。本実践は多面的な評価を行えるようデザインしている。

さて，評価について述べるならば，妥当性と信頼性について触れる必要がある。この点について次節で触れる。そして，そのあと，筆者らが実際にどのような取り組みや評価を行ったかや，評価の妥当性や信頼性についてどのように取り組んだかを述べる。

2 妥当性と信頼性

2.1 一般的な妥当性と信頼性

評価の妥当性と信頼性は大雑把にいうと，妥当性は評価しようとしていることが本当に評価できているかを検討するものであり，信頼性は評価・測定しようとしていることがらが複数時点，複数の評価者もしくは複数の項目の間で一致しているかどうかを検討するものである。

まず妥当性について。細かな議論は村山（2012）に任せるとして，妥当かどうかは，得られたデータの特性や構造をもとに検討する方法や，評価の内容や概念自体をデータに寄らず検討する方法などがある。

次に信頼性について。複数時点間での一致度は，同じテストか同一レベルの異なるテストの結果の一致具合をもとに検討する。複数の評価者間の方は，文字通り，同じ内容を複数の者で評価した結果の一致具合を検討する。複数の項目間の一致については，尺度の内的一貫性のことである。

これらの妥当性と信頼性は，評価・測定しようとする内容によって，検討されるものが異なったり，検討しようがないものがあったりする。ここで述べておくべきことは，妥当性と信頼性を等閑視することはできないということである。なぜなら，筆者らの取り組みのように，高校生を多面的に評価しようとするなら，評価者側が評価しようとしていることが本当に評価できているか，また，実施している評価方法は評価者が変わっても同様に評価できるかは，検討しておくべきことがらであり，落ち度がある場合は検討結果をもとに改善する必要があるからである。

では，筆者らがどのように評価の妥当性と信頼性を検討したかについて，次項で説明する。

2.2 筆者らが検討した妥当性と信頼性

2.2.1 取り組みの枠組みについて

筆者らが行っている課題探究の取り組みは，雨森ほか（2019）で説明しているが，再度説明しておこう。表１は雨森ほか（2019）でも用いた当日のスケジュールである。

本取り組みは，限られた材料を用いて，ある建造物をグループごとに制作すること（表１の⑤⑦）を中心に据え，設計・制作段階に入り前に，参加時点で不足している知識や考え方についての講義を受けたり（①），その確認の小テストや解説を受けたり（③④）するものである。また，グループの設計・制作内容等を参加者各自にプレゼンテーションすることを求めている（⑪）。プレゼンテーションの前に，８色の油性ペンを渡し，手書きでポスターを作成する時間を取っている（⑩）。

表１より，取り組みにおいて，小テストの採点，グループワーク時の高校生の参加状況の観察，ワークリストや最終レポートの内容，プレゼンテーションの内容とふるまいを評価していることがわかるが，これらのうち，本稿では，チェックリストとルーブリックについて取り上げる。

2.2.2 評価の方法

チェックリストは，グループワーク時に，高校生があることがらができているかを観察する際に用いたもので，評価の観点は，社会人基礎力をもとに作成した。具体的には，「物事に進んで取り組む力」，「他人に働きかけ巻き込む力」，「目的を設定し確実に行動する力」，「自分の意見をわかりやすく伝える力」，「相手の意見を丁寧に聴く力」，「意見の違いや立場の違いを理解する力」，「自分と周囲の人々や物事との関係性を理解する力」である。たとえば，「物事に進んで取り組む力」は，「やるべきことを自ら見つけて積極的に取り組んでいる」という行動ができているかを観察し，できていればリストの該当箇所をチェックすることになる。

ルーブリックは，バリュー・ルーブリックの口頭伝達力（Oral communication）をもとに作成している。なお，当ルーブリックの翻訳は松下（2012）に掲載されている。このルーブリックに，「説明内容（4 点）」「プレゼン時間」「質疑応答」を加えている。「説明内容（4 点）」とは，あらかじめプレゼン時に話すべき内容として指定した 4 つのポイントが，実際にきちんと話せているかを評価するものである [2]。

各評価者はおよそ 4 名の生徒の評価を担当し，各生徒は 2 名から評価を受ける体制をとっている。

2.2.3 評価の妥当性と信頼性の検討方法

まず妥当性について。チェックリスト，ルーブリッ

表１第２回目の実践

時間	内容	評価活動	使用プリント
① 9:40 ~ 10:15	講義１		
②10:15 ~ 12:30	予備実験 （試作と破壊実験） (昼食・休憩含む)		
③12:30 ~ 12:50	小テスト（提出）	採点（後日）	
④12:50 ~ 13:10	講義２		
⑤13:10 ~ 16:10	製作１（設計と製作）	観察	チェックリスト：設計、製作
⑥16:10 ~ 16:30	１日目ワークシート 完成（提出）	ワークシートを 評価（後日）	
◆2日目			
⑦ 9:30 ~ 12:30	製作２（昼食・休憩含む）	観察	チェックリスト：製作
⑧12:30 ~ 13:00	破壊試験		
⑨13:00 ~ 13:15	プレゼンテーションに 関する説明（TA）		
⑩13:15 ~ 14:25	ポスター作成 ＋ リハーサル(休憩含む)		
⑪14:25 ~ 15:00	プレゼンテーション	発表を評価	ルーブリック
⑫15:00 ~ 16:00	ワークシート， 最終レポート， アンケート（提出）	ワークシートと レポートを評価（後日）	
⑬16:00 ~ 16:30	まとめ，感想		

クとも，妥当性は，課題探求のあとに開催する，評価者間の反省会において検討している。そこでは，評価しようとしている項目が，当日の内容できちんと評価できているか，できないならどうすれば評価できるかを検討している。

次に信頼性について。チェックリストでは，該当する行動がとれていかどうかを，1 名の生徒に対して，2 名で評価している [3]。評価結果を評価項目ごとに一致率とκ係数を求め，信頼性を検討する。

ルーブリックも，1 名の生徒に対して，複数名で評価している。こちらも一致率とκ係数で信頼性を検討する。

ところで，κ係数は，教科書的な説明をすれば，評価が名義尺度なら単純なκ係数，順序尺度なら重み付きκ係数を用いることになっている。両者の大きな違いは，評価が一致していない場合のズレの幅を考慮するかどうかであり，前者は考慮しない，後者は考慮するものである。後者で，評価者間のズレが大きければ前者よりκ係数は小さく，ズレが小さければ前者よりκ係数は大きくなる。すなわち，ズレの幅次第で，単純なκ係数より，重み付きのほうが，数値が低くも高くも，すなわち，信頼性が低くも高くもなるわけである。

尺度の水準だけで考えれば，本稿の場合，重み付きκ係数を求めることになる。しかし，完成度の高いルーブリックを用いた評価において，4 段階で言うところの 1 と 4 のように評価が大きく割れることはあまり起こりにくい。評価が割れにくい場合，重み付きκ係数を求めると，簡単なものよりも甘くなる傾向になる。

また，課題探究の取り組みでの評価を入試に取り込むことを考えると，ズレの幅ではなく，一致しているかどうか自体が重要である。これらのことを勘案し，本稿では単純なκ係数を求めている [4]。

3 評価の妥当性と信頼性の検討

3.1 妥当性の検討

課題探究の当日からおよそ 1 ヵ月後に評価者が集まる反省会を開催し，評価の妥当性などを議論している。反省会には，各評価の基本的な分布に関するデータを示しながら，議論を行っている。具体的な検討内容を次に示す。

チェックリストについて。チェックリストでは，ある行動が見られたかどうかを，「見られなかった」，「1 度は見られた」，「数度見られた」のように頻度も併せて評価している。たとえば，「物事に進んで取り組む力」は，表 2 のような形式で評価している。

反省会において，この頻度について，「ゼロ回と 1 回以上は判別できるが，1 回と数度を判別する自信は持てない」，「行動に移せるかどうかを評価するべきであって，頻度よりも，ゼロ回かどうかを評価するべきでは」などの意見が寄せられた。これらの意見をもとに，頻度を問うことを辞め，行動が「見られた」かどうかをチェックするように改めることにした。

このほか，「似通った観点があるので整理が必要」，「観点に対応する場面が現れなかった」，「単純に観点が多いので評価しきれない」などがあり，観点の整理を行うようにした。

発表に関するルーブリックについては，概ね評価しやすいという意見であり，妥当な内容になっているということであった。ただ，発表時間を守れているかは，評価者各自がするのではなく，各グループにいる評価

表 2 チェックリストの例

評価項目、行動例	生徒	行動	数度見られた	1 度は見られた	見られなかった	備考（気になるネガティブな行動等）
【主体性】		①				
		②				
物事に進んで取り組む力		①				
①やるべきことを自ら見つけて積極的に取り組んでいる		②				
②他人が嫌がることにも自ら取り組んでいる		①				
		②				
		①				
		②				

者のうち1名が行えばいいのではとの指摘があり，次回以降そのように変更した。

3.2 信頼性の検討

3.2.1 チェックリストの信頼性

チェックリストの評価の場面は，先ほど示した表1の「製作1」と「製作2」である。「製作1」と「製作2」では，生徒たちは，各班で製作する物の設計図を描き,設計図をもとに製作する。作業の性質上,「製作1」では，生徒たちの動きがあまり見られなかったこともあり，以下では比較的動きが見られた「製作2」での評価について検討する。

表3は「製作2」で行われた評価に対して信頼性を確認するものである。表には，評価者間の単純な一致率と，単純な一致率から偶然の一致を考慮したκ係数を示している。両者とも1に近いほど評価が一致している，すなわち信頼性が高いことを意味する。なお，当然のことながら単純な一致率はゼロを下回ることはないが，κ係数はそうなることはあり得る。両者の関係は，計算上，単純な一致率がκ係数を下回ることはなく，信頼性の指標として数値が低い目，すなわち信頼性が低い目になるのはκ係数のほうである[5)]。両者の値を確認してみよう。

まず，単純な一致率について。一致率の良し悪しを決めることに基準はないが，1/2が不一致なのは感覚的に良くないだろう。もう少し厳しく見ても，1/3がズレているのも良くないと思われる。ここでは，評価の不一致は1/3まで許容するとする。その結果,「目的を設定し確実に行動する力」の①,「相手の意見を丁寧に聴く力」の①と②,「意見の違いや立場の違いを理解する力」,「自分と周囲の人々や物事との関係性を理解する力」の②の5つが基準値を満たさない。残り8項目についてκ係数を確認してみる。

κ係数の良し悪しの基準は，Landis and Koch (1977) によると，ゼロ未満で不十分，ゼロから0.2間隔で，ほんの少し一致，一応一致，適度な一致，概ね一致，ほぼ一致となっている。これらのうち，一応一致（0.41 ～ 0.6）以上の基準で，先ほどの残りの8つについてみてみると,「物事に進んで取り組む力」の②,「他人に働きかけ巻き込む力」の②の2つが基準値を満たさないことになる。

単純な一致率，κ係数を確認した結果，6つの項目は信頼性がある程度高いということがわかった。ところで，信頼性が高い，言い換えるなら評価者間の一致度が高いとしても，評価しようとしている行動が，評価場面においてほとんど生起しない状況であるなら，評価が一致していても，評価自体意味をなさないことになる。よって，最後に，残り6項目について，評価対象の行動が見られたかどうかを表す行動未確認率をもとに吟味する。

表3「製作2」の評価の信頼性

		一致率	κ係数	行動未確認率
【主体性】 物事に進んで取り組む力	①	0.83	0.57	2.8
	②	0.75	0.31	72.2
【働きかけ力】 他人に働きかけ巻き込む力	①	0.75	0.48	69.4
	②	0.67	0.37	58.3
【実行力】 目的を設定し確実に行動する力	①	0.50	0.13	58.3
	②	0.75	0.44	63.9
【発信力】 自分の意見をわかりやすく伝える力	①	0.92	0.81	58.3
	②	0.92	0.76	83.3
【傾聴力】 相手の意見を丁寧に聴く力	①	0.50	0.17	36.1
	②	0.58	0.12	66.7
【柔軟性】 意見の違いや立場の違いを理解する力		0.58	0.29	61.1
【状況把握力】 自分と周囲の人々や物事との関係性を理解する力	①	0.67	0.47	25.0
	②	0.50	0.17	63.9

この行動未確認率は，評価の対象となる行動が見られなかったと評価した総数を全数で除したもので，100 に近ければ行動が見られなかったこと，0 に近ければ行動自体は見られたことを意味する。この値を見ると，自分の意見をわかりやすく伝える力」の②は 83.3 と高く，明らかに行動自体が生起していないことがわかる。残りの 5 項目については，行動が現れにくいか，評価自体が適していないかを検討する必要がある。これは妥当性の点でも検討課題となる。

この行動未確認率を見ると，信頼性の点では，「相手の意見を丁寧に聴く力」の①は基準値以下であったが，評価の場面において行動は起きていたことがわかる。この項目については評価の内容を変え，信頼性を高められれば，評価し得るものになる可能性を秘めている。

3.2.2 ルーブリックの信頼性

ルーブリックについても一致率と κ 係数を確認し，信頼性を検討する（表 4）。こちらは，チェックリストの方とは異なり，行動が起きないことはありえないので，行動未確認率は示していない。

チェックリストと同様に，一致率については不一致を 1/3 まで許容，κ 係数については一応一致（0.41 ～ 0.6）以上の基準で確認すると,「言葉の選び方」,「ポスターの工夫」が基準値を満たしていない。

妥当性を検討した反省会では，評価者はこの 2 項目とも評価しやすいという感触を持っていたが，ルーブリックで用いられている言葉の共通理解が図られておらず，評価者間で一致しにくいことになったのだろう。この 2 項目については，事前に打ち合わせや研修を行い，共通理解を図る必要がある。

3.3 チェックリストのさらなる検討

本節では，チェックリストとルーブリックの妥当性と信頼性について述べてきた。妥当性については，課題探究実施後に行う反省会に基づいて，信頼性については κ 係数等を用いて，議論を進めてきた。

表 4 「プレゼンテーション」の評価の信頼性

	一致率	κ 係数
話し方	0.75	0.43
言葉の選び方	0.58	0.34
ポスターの工夫	0.50	0.22
説明内容 4 点	1.00	1.00
プレゼン時間（5 分）	0.75	0.48
質疑応答	0.83	0.63

ルーブリックの信頼性は，事前の打ち合わせや研修によって数値を高められると予想される。しかし，チェックリストの信頼性は，信頼性の低さが場面に依存するのか，項目自体に依存しているのかを議論しないと，妥当性も怪しくなると考えられる。チェックリストの各項目について，さらに検討する。全項目をつぶさに検討すべきであることは承知しているが，紙幅の都合もあり,「物事に進んで取り組む力」の①と②,「相手の意見を丁寧に聴く力」の①と②の 4 項目について限って検討する。

まず,「物事に進んで取り組む力」であるが，具体的には，①は単純に積極的な行動ができているかを，②は他人が嫌がることについて積極的に行動できているかを評価している。つまり，①ができないと②はできない。また，②は他人が嫌がることかどうかが評価者で判断しづらい。これらのことが，②の信頼性が低くなった原因であると筆者らは考えている。

次に「相手の意見を丁寧に聴く力」の 2 項目について。この 2 項目とも κ 係数が低くなっているが，原因は場面の設定と評価項目の対応の悪さであると筆者らは考えている。前述した通り，チェックリストは，限られた材料で，ある建造物をグループごとに制作する場面で用いられている。とくに「製作 2」は，設計図を描き終わったあとの製作段階のグループワークであることから，意見交換が行われるよりも，設計図をもとに作業が進められる。このこともあり，会話を前提とした行動は生起しにくく，明確に現れていない行動を甘く評価する者と厳密に評価する者で評価が割れ，信頼性が低くなっていると考えられる。実際に反省会では，当項目に限らず，どの程度の行動であるならば，行動が見られたとしていいか悩ましいという声もあった。

4 おわりに

本稿では，筆者らが行っている課題探究の取り組みでの評価方法について，本文ならびに注の中で説明してきた。

筆者らが用いているチェックリストやルーブルックは，信頼性を検討する過程で完全ではないことはわかっている。最初から完全なものを作ることは不可能に近く，筆者らも次の実践で改善を図るために，反省会等を行っているわけである。また，評価慣れや事前の打ち合わせ不足が評価の信頼性を低めている可能性も示唆された。知識・技能以外の主体性等を評

価するには，評価者側の事前の準備が肝要である。

大学は高大接続改革を受け，入試における学力の多面的評価について，高校は学習指導要領の改訂を受け，探究活動の実施ならびにその評価について悩んでいるのが現状であろう。筆者らは双方の悩みを，双方の関係者が協働して取り組み，研究しながら解消しようとしている。もちろん，参照し，役立つ情報は，書籍やインターネット上などに多くあるが，何が正しくて，適切なのかはわかりにくい。たとえば，本稿では，評価の場面において，チェックリストを用いていたり，信頼性の検討において重み付きκ係数ではなく単純なκ係数を求めたりしているのは，教科書や通説を妄信的に従わずに，内容を吟味した結果である。ただし，筆者らの方法が適切かどうかは，まだ実践や議論が必要である。

筆者らは，本稿で触れた取り組み以降，改善の過程を経て，さらに実践を積み重ねている。この改善された実践については，別の機会に紹介したい。

注

1) 本研究を進めるにあたり，静岡県教育委員会，静岡県私学協会，各校の学校長から承諾を得ている。

2) 筆者らは，雨森ほか（2019）において「独自の○○力を設定する必要性があるなら，設定するべきだと考えるが，大学入試で評価する能力はそこまでオリジナリティが求められるものではなく，結果的に独自の○○力は，経済産業省が提唱する社会人基礎力や国立教育政策研究所が提案する21世紀型能力，OECDキー・コンピテンシーなどで代替可能だと考えられる。他に参照可能な尺度等があるならば，参照したほうが，開発のコストを抑えられたり，妥当性や汎用性を高められたりなど，メリットがある。」と述べたように，むやみに○○力を設定することを避けたり，汎用性を志向したりしている。

3) 課題探求型の取り組みにおいて，なんでもルーブリックを用いて評価しようとする嫌いがあるが，筆者らはルーブリックを用いるべきかどうか自体を検討する必要があると考えており，検討の結果，チェックリストによる評価を行うことにした。

4) κ係数について，単純なもの，重み付きのもの，どちらを算出するかを議論するくらいなら，両者を算出して，値と評価の分布をもとに検討するほうが建設的だと筆者らは考えているが，ここであえて述べようとしているのは，通例だから，または，尺度水準が○○だから妄信的に重み付きのほうを算出することへの危うさを感じているからである。教科書，慣例，表計算ソフトや統計ソフトのみに頼るのではなく，研究者自身で適切な方法を選択するようにしたいと筆者らは考えている。

5) 本稿のような評価の信頼性を検討する際に，単純な一致率で十分か，κ係数を用いるべきか，はたまた，他の指標が妥当であるかは，もう少し検討を重ねたい。個人的には，単純な一致率と分布の確認で十分ではないかと考えている。

参考文献

AAC&U Value
< https://www.aacu.org/value>（2018年3月16日）

雨森聡・宇佐美壽英・藤井朋之 (2019).「パフォーマンス課題を用いた主体性等を評価するデザイン――静岡県における工学系の高大接続事例をもとに」『大学入試研究ジャーナル』**29**, 188-193.

中央教育審議会 (2014).「新しい時代にふさわしい高大接続の実現に向けた高等学校教育，大学教育，大学入学者選抜の一体的改革について（答申）」

中央教育審議会 (2016).「幼稚園、小学校、中学校、高等学校及び特別支援学校の学習指導要領等の改善及び必要な方策等について（答申）」

Landis JR, Koch GG.(1977). “The measurement of observer agreement for categorical data”, Biometrics. 33(1):159-174.

松下佳代 (2012).「パフォーマンス評価による学習の質の評価：学習評価の構図の分析にもとづいて」『京都大学高等教育研究』8, 75-114.

村山航 (2012).「妥当性概念の歴史的変遷と心理測定学的観点からの考察」『教育心理学年報』**51**, 118-130.