項目分析：心理尺度の信頼性と妥当性

信頼性とは測定の精度を意味し，妥当性とは測定された内容が意図したものとどれだけ一致しているかを意味している。ここでは，クロンバックのα係数（Cronbach's coefficient alpha）の計算方法を中心に，心理尺度の精度と妥当性を高める方法にどのようなものがあるかを説明する。

信頼性の検証

信頼性の高い尺度とは，その尺度の個別的な項目間で一貫性があると考えられる。たとえば，ある質問項目で Yes と回答した被験者は，同じ尺度内の別項目でも Yes と回答するはずである。このような一貫性をチェックする尺度として，クロンバックのα係数があるが，これ以外にも以下のような様々な分析手法が存在する。

G-P 分析（Good－Poor Analysis）

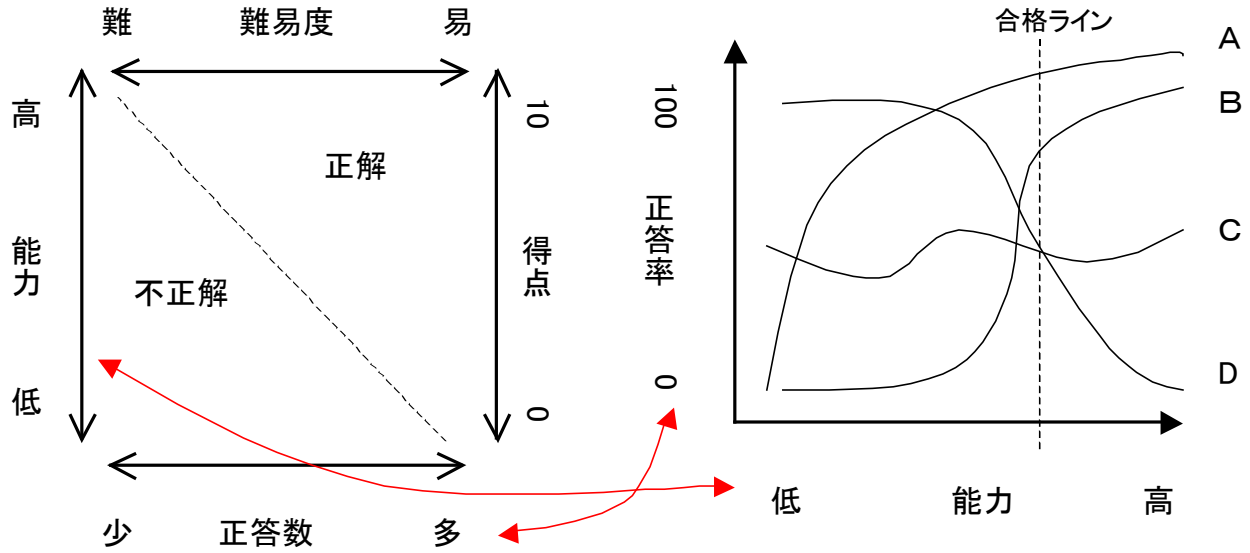
たとえば，合計得点の高低によって被験者を分割する。どの項目についてもその平均値は上位群の方が下位群より高いことが予想される（逆転項目などは適宜変換しておく）が，高低双方のグループの平均値を計算し，グループ間で平均値の差の検定（e.g., t 検定）を行えば有意な差が見られるはずである。理想的には，上位軍，下位群で差が見られた項目のみを残して，それ以外は尺度から取り除く操作も考えられ得るが，相対的な差の大きさによって判断するのが普通である。

I-T 相関分析（Item-Total Correlation Analysis）

尺度得点の高い被験者は，それぞれの項目でも高いと予想される。したがって，項目得点と尺度得点との相関係数を見て，あまりにも低い項目は尺度から取り除く。

S-P 表（Student-Problem Table）を用いた分析

S-P 表は主に成績評価の関係で用いられることが多いが，この分析は一般的な心理尺度のチェックにも適用可能である。たとえば，以下の左図のように，被験者を 10 点満点の尺度の上位群と下位群とに分け両群の回答パターンを比較することが一般的であるが，能力や得点を何か別の尺度得点に置き換えて考えることも可能である。



上図左に示したとおり，誤答は右下のエリアに集中するはずである。しかし，問題によってはできる生徒ほど解けない問題というのも存在するかも知れない。この関係は右図に示したような項目特性曲線を描くとはっきりする。たとえば，A～D の問題のうち，どの問題が入学試験問題に相応しいかはすぐに分かるであろう。

折半法（Split-half correlations）

尺度全体を同等と見なすことのできる 2 つの尺度に折半し，それぞれの観測値（尺度得点）間の相関係数を求めれば，これを信頼性の指標とすることもできるであろう。同様の発想に基づいた検討方法としては再検査法（Test-retest correlations）がある。

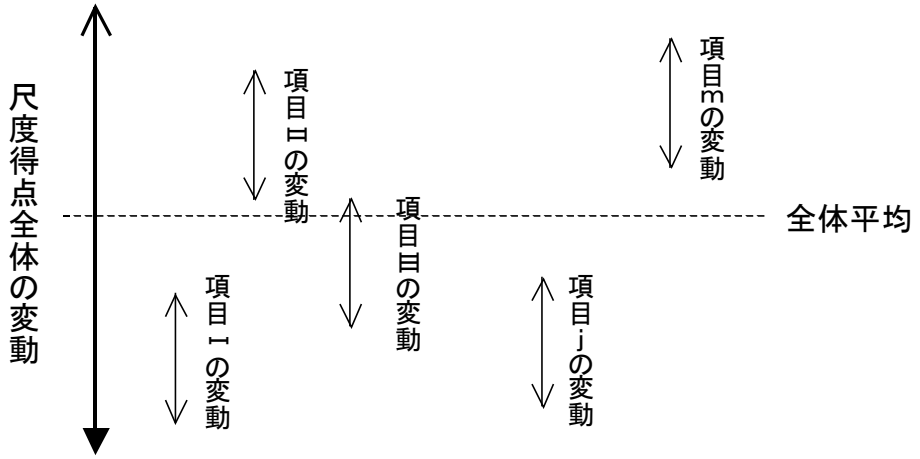
クロンバックのα係数（Cronbach's coefficient alpha）

上記の折半法には，折半する方法が 1 通りではないという問題がある。すなわち，可能な全ての折半方法を考慮した信頼性の推定値を求めた方が適切であることが分かる。このような推定値がクロンバックのα係数である（単にα係数と呼ばれることが多い）。以下の式に基づいて計算する。

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\sum_{j=1}^{m}\sigma_j^2}{\sigma_x^2}\right) = \frac{m}{m-1}\left(\frac{\sum_{i=1 i\neq j,}^{m}\sum_{j=1}^{m} S_{ij}}{\sum_{j=1}^{m} S_j^2 + \sum_{i=1 i\neq j,}^{m}\sum_{j=1}^{m} S_{ij}}\right)$$

このとき，mをその尺度の中の質問項目の数，σ_xを尺度得点全体の分散，σ_jを各質問項目の分散とする。また，S_jを項目 j の分散，S_{ij}を共分散（これを標準化すれば相関係数）とする。なお，$\sigma_x^2 = \sum_{j=1}^{m} S_j^2 + \sum_{i=1 i\neq j,}^{m}\sum_{j=1}^{m} S_{ij}$ である。

本来であれば，$x_i = t_i + e_i$，$\sigma^2(x) = \sigma^2(t) + \sigma^2(e)$（$x_i$は観測値，$t_i$は真の値，$e_i$は測定誤差）のとき，$\rho = \frac{\sigma_t^2}{\sigma_x^2}$ を信頼係数とする・・・といったところから導出過程を説明すべきであるが [1]，ひとまず，尺度の一貫性を分析する分には上記の式を使いこなせれば十分であろう。α係数の値が 1 に近づくほど一貫性が高い尺度といえるので，一般的な利用上の目安としては，0.7～0.8 を目指して各下位項目を付けたり外したりしていくことになる。



さて，初めの式を見れば分かるとおり，質問項目mの数が大きくなるほど尺度の信頼度が高まる。これは 1 つの質問よりも 2 つの質問の方が好ましく，そして 2 つよりも更に多くの質問をした方がよく調べることができるという直感的な理解とも合致しているといえるであろう。尺度全体の分散と質問項目得点の分散合計の比も，上記の図のように考えるとそれほど突飛な話ではないことが分かるであろう。たとえば，項目どうしの反応パターンに一貫性があれば，$\sum_{j=1}^{m}\sigma_j^2$ の値はσ_xよりも小さくなるはずである。これは，一部の質問項目を逆転項目に変更した場合

[1] 一要因の分散分析の証明を参照するとその雰囲気がなんとなく分かると思う。

などをイメージすればよく分かると思う。逆転項目に変更した場合には全体の変動σ_xは大きくなるが，$\sum_{j=1}^{m}\sigma_j^2$の値には変化がないはずである（タウ等価性の直感的理解)。また，最後の式を見ると，項目間の相関係数が高いほど信頼性が高くなることが分かる。つまり，相互に相関が高い項目を選び，項目数を増やすことによって尺度の一貫性（内的整合性：internal consistency）が高くなることが分かるであろうが，これも比較的直観とうまく適合するように思われる。

妥当性の問題

上述した様々な分析によって尺度の一貫性，信頼性を高めることができる。たとえば，I-T 相関分析を行ったり，尺度内の各下位項目を付けたり外したりしてα係数が高くなる条件を探ることもできるし，S-P 表を作成して識別率の低い問題をテストから除外することもできるようになる。

しかし，信頼性が高いことと，妥当な尺度であることとは別物である。たとえば，測定しようとしている概念と尺度とが論理的に対応しているかどうかは保証されていない。つまり，信頼性を高めようと努力した結果，尺度が特定の内容に偏った質問項目ばかりになってしまうこともあり得るわけである。このような問題は論理的妥当性の問題，内容的妥当性の問題と呼ばれている。

そして，心理尺度を作成する目的，質問紙を作成する目的を考えると，更に別の妥当性の問題もでてくることがわかるであろう。たとえば，何か新しい心理尺度を作成するときには，普通は既に存在する別の尺度との関連性を調べたい場合がほとんどであろう。また，ある特殊な集団と別の集団との識別を容易にするための尺度作成というのも，重要な実用的な意味を持つのでよく見られる。このような研究状況で問題とされる妥当性として，基準関連妥当性が存在する。

この妥当性は，新たに作成した心理尺度と既存の心理テストの相関を調べたり，何か別の独立変数を設定して G-P 分析を行ったりすることによって検証される。たとえば，学習方略に関する尺度を作成したならば，学業成績の高低で被験者を分割し，優秀者と劣等者を独立変数として各項目得点に差が見られるかをチェックする必要もあるだろう（具体的には以下の表を参照)。

	項目得点（高い）	項目得点（低い）
成績（優秀者）	30	3
成績（劣等者）	8	20

（クロス集計表に対してχ二乗検定を行うことができる）

なお，上記の妥当性に関するチェック項目（論理的妥当性，基準関連妥当性）は項目分析以外にも，尺度得点全体についても検証が必要とされる項目であるのはいうまでもない。

たとえば，探索的な因子分析を行うと，どんなにいい加減に作った質問紙でも最低 1 つはもっともらしい因子が発見（捏造？）される。しかし，当然のことながらその因子が本当に実在するかどうかの保証は全くない。また，いくら真面目に作成した質問紙によって測られても，その因子が本当に実在すると言っても良いかどうかは，その後の追加研究によって検証をまたなければならないのが普通である。そして，その因子の実在性やもっともらしさを検証する方法として，たとえば，特定の因子得点を従属変数とした分散分析やχ二乗分析などがしばしば実施されるのである。