K-036

eテスティングにおける得点・時間予測システムの開発

Development of Prediction System of Score and Time in e-Testing

ソンムァン　ポクポン†　植野　真臣†
Pokpong Songmuang†, Maomi Ueno†

1．はじめに

近年，e テスティングの普及に伴い，アイテム・バンク方式のテスト構成が一般化して，テスト得点および時間の予測の実現が可能になりつつある．例えば，永岡(1)が先駆的に切断指数分布を用いた予測得点分布を提示するテスト構成支援システムを開発している．Ueno(2) は，インターネット上における e テスティングのアイテム・バンク統合システムを開発し，二項分布，混合二項分布を予測得点分布とするウェブ・ベース・テスト構成支援システムを開発している．一方，テスト得点分布のみに焦点化した研究は，Keats and Lord(3) がテスト得点分布としてベータ二項分布を当てはめていた研究が知られているが，その他にはほとんど存在しない．

それに対し，テストの所要時間分布に関する研究は多い．教育工学分野では，レスポンス・アナライザーやコンピュータ・テストより得られる所要時間データの解析手法として集団学習応答曲線（Response Curve) (例えば，(4)(5)) は古くから研究されている．また，計量心理学分野でも，Thissen(6) が対数正規分布モデルのテスト所要時間データへの当てはめを行っている．また，Verhelst ら(7) は心理テストにおけるスピードテストにガンマ分布を用い，Roskam(8) はテスト所要時間にワイブル分布を当てはめている．

以上のレビューより明らかなように，一部のモデルについて当てはまりの比較研究はされているものの，テストの得点・時間の予測という観点では，ほとんど比較研究がおこなわれていない．そこで，本論では，まず，実データを用いたテスト得点および所要時間についての予測モデルの比較実験を行う．ただし，テスト得点予測モデルはこれまでほとんど研究されてこなかったため，本論では，1) ベータ二項分布，混合二項分布を組み合わせて拡張した混合ベータ二項分布，2) 項目反応理論（ラッシュ・モデル，2 パラメータ・ロジスティック・モデル（2PL））を用いたテスト得点分布を提案し，従来に用いられてきた二項分布，混合二項分布，ベータ二項分布，切断指数分布と比較することにする．また，テスト所要時間予測モデルとしては，これまで提案されてきた分布（正規分布，対数正規分布，拡張ガンマ分布，ワイブル分布）による比較を行う．

実験の結果，テスト得点予測モデルとして項目反応理論（2PL），テスト所要時間予測モデルとして拡張ガンマ分布モデルが最も良い予測精度を示した．

以上の結果を用いて，テスト構成過程においてテストの予測得点分布，予測所要時間分布の状態を逐次可視化するウェブ・ベース・テスト構成支援システムを提案する．さらに，システム評価実験が行い，本システムの有効性を示した．

† 電気通信大学大学院　情報システム研究科

2. e テスティング・システム

著者らは，これまで Web 上で動作する統合型 e テスティング・システムを開発してきた(9)．システムは，図 1 に示すように以下のモジュールによって構成されている．1）項目作成支援システム 2)アイテム・バンク 3)テスト実施システム 4)テスト構成支援システム 5)テスト・データベース 6)データ分析システム．

3. 予測モデルの比較実験

3.1 実験方法

本章では，e テスティングにおけるテスト構成支援システムのためのテスト得点・時間予測システムを構築するために，様々な得点分布と所要時間分布を実データを用いて評価する．本実験で用いるデータは，表 1 に示すような N 大学工学部の実際の授業の中で実施された 4 つのテスト・データである．具体的には，これらのデータを用いて，1) 得点分布，所要時間分布の実データへの当てはまり評価，2) 新たに構成されたテストの予測得点分布，予測所要時間分布の予測精度評価，を行う．特に 2) では，データ中，50% のデータをランダム・サンプリングしてトレーニング・データとし，残りの 50% のデータをバリデーション・データとして，トレーニング・データより推定された予測分布とバリデーション・データとの当てはまりを評価し，これを 1000 回繰り返すことにする．ただし，実験 1)，2)において，得点分布の評価は得点分布，実データともに 10 段階の離散分布に変形し，所要時間分布の評価は，所要時間分布，実データともに累積分布に変形して，分布と実データとの二乗平均平方根誤差（RMSE: Root Mean-Square Error）を計算することにより評価した．

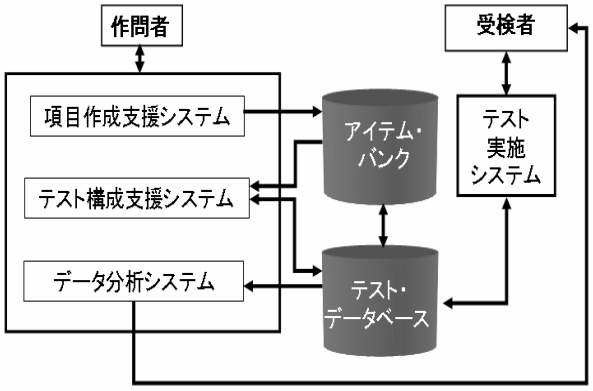

図 1 e テスティング・システムの構成

表 1 サンプルデータの要約

	テスト内容	項目数	受検者数	正答率	標準偏差	項目平均所要時間(秒)	標準偏差
A	日本語能力試験 3 級	69	91	0.445	0.198	1421.05	1366.60
B	コンピュータ初級	54	62	0.270	0.217	659.10	569.54
C	情報処理	20	72	0.739	0.903	204.11	284.33
D	ネットワーク技術	37	66	0.241	0.142	622.91	614.07

表 2 得点分布と実データおよび予測得点分布と実データの RMSE

得点分布と実データの RMSE	二項分布	混合二項分布	ベータ二項分布	混合ベータ二項分布	ラッシュ・モデル	2PL	切断指数分
A	0.00356	0.00042	0.00052	0.00041	0.00059	0.00041	0.00052
B	0.00570	0.00235	0.00319	0.00120	0.00160	0.00118	0.00159
C	0.00766	0.00332	0.00321	0.00262	0.00267	0.00192	0.00277
D	0.00464	0.00098	0.00178	0.00060	0.00093	0.00075	0.00079
得点分布と実データの RMSE	二項分布	混合二項分布	ベータ二項分布	混合ベータ二項分布	ラッシュ・モデル	2PL	切断指数分
A	0.00346	0.00056	0.00062	0.00055	0.00067	0.00053	0.00062
B	0.00591	0.00239	0.00324	0.00134	0.00168	0.00132	0.00173
C	0.00737	0.00339	0.00331	0.00269	0.00279	0.00179	0.00286
D	0.00422	0.00105	0.00182	0.00073	0.00095	0.00068	0.00089

表 3 所要時間分布と実データおよび予測所要時間分布と実データの RMSE

所要時間分布と実データの RMSE	正規分布	対数正規分布	拡張ガンマ分布	ワイブル分布
A	0.758	1.195	0.157	0.920
B	1.043	3.167	2.812	4.266
C	5.232	1.226	3.199	1.271
D	0.979	1.294	0.534	1.330
所要時間分布と実データの RMSE	正規分布	対数正規分布	拡張ガンマ分布	ワイブル分布
A	4.481	1.961	1.032	1.802
B	4.261	4.455	1.578	4.271
C	4.826	2.954	1.493	1.630
D	1.007	2.080	1.002	1.220

3.2 得点分布における実験結果

まず本節では、新得点予測モデルを提案される．ベータ二項分布，混合二項分布を組み合わせて拡張した混合ベータ二項分布は以下のように推定される．

$$p(x|\alpha,\beta)=\sum_{i=1}^{m}\left[\frac{1}{m}\binom{m}{x}\frac{B(\alpha_i+x,\beta_i+n-x)}{B(\alpha_i,\beta_i)}\right] \quad (1)$$

ここで，$i,(1,\ldots,m)$ は i 番目の項目を示し，n は受検者の総数である，$B(\alpha,\beta)$ はベータ関数を示し，α と β は項目の推定されたパラメータである．

項目反応理論は様々なモデルが提案されているが，その中でも最も良く用いられるラッシュ・モデル(10)と 2PL(11)の二つを用いることにする．項目反応理論では，項目u_{ij}に対して受検者$j,(1,\ldots,n)$の反応$i,(1,\ldots,m)$を以下のように定義する．

$$u_{ij}=\begin{cases}1:\text{学習者jが項目iに正答したとき}\\0:\text{上記以外}\end{cases}$$

2PL の正答する確率は以下のように定義する．

$$p(u_{ij}=1|\theta_j)=\frac{1}{1+\exp\{-1.7a_i(\theta_j-b_i)\}} \quad (2)$$

ここで，θ_jは受検者jの能力パラメータを示し，b_iは項目 i の難易度パラメータを示し，a_i は識別パラメータを示す．ラッシュ・モデルは 2PLのa_i=1 場合のモデルである．また，両モデルともにθjの分布は標準正規分布を仮定している．

ここでは，項目反応理論を用いたテスト得点分布モデルを以下のように定義する．

$$p(x)=\int_{-\infty}^{\infty}\sum_{i=1}^{m}\frac{1}{m}\binom{m}{x}p(u_{ij}|\theta_j)d\theta_j \quad (3)$$

次に，表 1 の実データから得点分布モデル，二項分布，混合二項分布，ベータ二項分布，混合ベータ・二項分布，項目反応理論（ラッシュ・モデル，2PL），切断指数分布のパラメータを推定し，推定された分布と実データとの当てはまりを前述の RMSE によって評価する．

推定された分布と実データとの誤差を示す RMSE は，表 2 の上半部の「得点分布と実データの RMSE」の欄にそれぞれ 4 つのテストの結果を示し，すべての得点分布モデルのうち，最も誤差の小さな得点分布モデルの RMSE の数値に下線を付して示した．

結果より，テスト A，B，C のデータでは項目反応理論における 2PL を用いたテスト得点分布が最も当てはまりが良く，テスト D のデータでも混合ベータ二項分布が最も良い当てはまりを示した．また，すべてのデータで当てはまりの良い上位 2 位は本論で提案した得点分布であった．

次に，前述の実験方法を用いてそれぞれの得点分布モデルを予測得点分布として用いた場合の予測精度を評価した．

推定された予測得点分布と予測された実データとの誤差を示す RMSE は，表 2 の下半部の「予測得点分布と実データの RMSE」の結果を示した．結果より，テスト A，B，C，D のデータで項目反応理論における 2PL を用いたテスト得点分布が最も高い予測精度を示した．

以上より，テスト得点分布として，当てはまりおよび予測精度の観点から見て，項目反応理論における 2PL が最も良いことが示され，主目的である e テスティングにおける得点分布の予測システムに採用するに適していると考えられる．

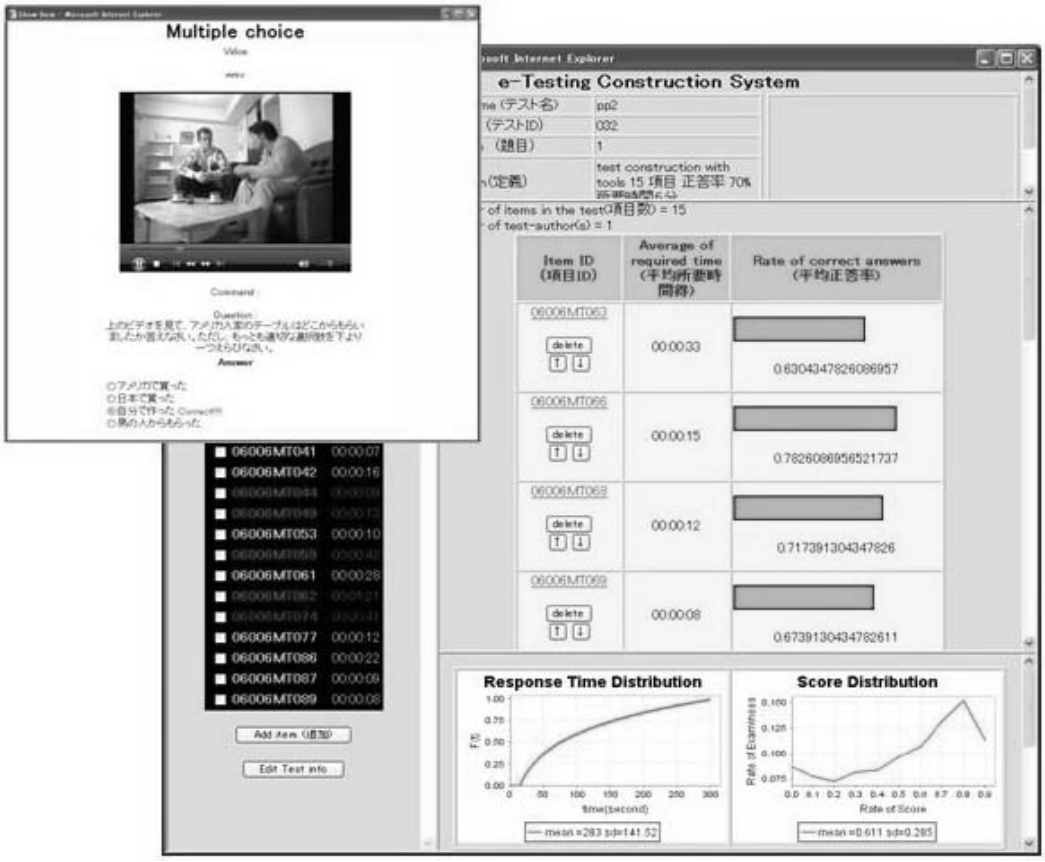

図 2 e テスティング構成支援システム

3.3 所要時間分布における実験結果

本節では，表 1 の実データから，これまで提案されてきた主な所要時間分布である正規分布，対数正規分布，拡張ガンマ分布，ワイブル分布のパラメータを推定し，推定された分布と実データとの当てはまりを前述の RMSE によって評価する．

推定された分布と実データとの誤差を示す RMSE は，表 3 の上半部の「所要時間分布と実データの RMSE」の欄にそれぞれ 4 つのテストの結果を示し，すべての所要時間分布モデルのうち，最も誤差の小さな得点分布モデルの RMSE の数値に下線を付して示した．結果より，テスト A と D では拡張ガンマ分布が最も当てはまりがよく，テスト B では正規分布が最も当てはまりがよく，テスト C では対数正規分布が最も当てはまりがよい．拡張ガンマ分布が最も成績が良いが，ばらつきが大きく，テスト所要時間分布の良さを決定するまでの結果ではないともいえる．

次に，前述の実験方法を用いてそれぞれの所要時間分布モデルを予測所要時間分布として用いた場合の予測精度を評価した．

推定された累積予測所要時間分布と予測された実データとの誤差を示す RMSE は，表 2 の下半部の「予測所要時間分布と実データの RMSE」の結果を示した．結果より，テスト A，B，C，D のデータで拡張ガンマ分布が最も高い予測精度を示した．

以上より，所要時間分布は予測精度の観点から見て，拡張ガンマ分布が最良で，主目的である e テスティングにおける得点分布の予測システムに採用するに適していると判断できる．

4. e テスティング構成支援システム

前章において，予測得点分布，予測所要時間分布の予測モデルには，それぞれ，項目反応理論（2PL）を用いた得点分布，拡張ガンマ分布による所要時間分布が適していることを示した．ここでは，著者らが開発してきた e テスティング構成支援システムに予測分布の提示機能を実装する．

図 2 が開発されたシステムのインターフェースである．インターフェース最上部には，テスト構成者が入力したテスト属性（テスト名，テストの ID，テスト構成プロセス ID，その他）が提示されている．また，画面左に提示されている ID は項目 ID である．項目 ID 上にマウスをあわせるとその項目内容が提示され，さらにクリックすると図のように項目内容が別ウインドウに提示される．複数の項目内容を同時に重ね合わせて提示することができ，書類を扱うようにテスト構成ができる．次に，テストに出題する項目 ID を選択し，項目 ID の左部にあるチェックボックスをクリックする（印を入れる）ことにより，インターフェース画面の中央にあるテスト・メモリに当該項目が登録される．「Delete」ボタンを押すことにより，その項目のテスト・メモリへの登録がキャンセルされる．テスト・メモリに登録された項目によって構成されるテストの予測テスト

得点分布および予測所要時間分布が計算され，画面最下部に提示される．画面左のグラフが予測所要時間分布，画面右のグラフが予測得点分布である．また，それぞれのグラフの下には，予測平均得点および予測得点標準偏差，予測平均所要時間および予測所要時間標準偏差の値が提示される．テスト構成者はこの予測システムを用いて，テスト構成を打ち切るか，項目を入れ替えねばならないか，新たに項目追加が必要か，といった意思決定を行うことができる．

5. システム評価

5.1 予測評価

本章では，本論で開発したｅテスティング構成支援システムを評価する．ここでの評価方法は，構成したいテストの目標とする平均得点とその標準偏差及び平均所要時間とその標準偏差をあらかじめ決めておき，本システムおよび比較用システムを用いて構成されたテストの目標との誤差を比較する．具体的には，本論で提案するシステムを「システム Ⓐ」，システム Ⓐ からインターフェース最下部に示された予測機能を除いたシステムを「システム Ⓑ」，さらにシステム Ⓑ からインターフェース中央に表示される項目の正答率，所要時間情報を表示しないシステムを「システム Ⓒ」として比較評価する．被験者として N 大学の日本人大学院生 30 名をランダムに 10 名ずつシステム Ⓐ, Ⓑ, Ⓒ に分配し，それぞれのシステムでテスト構成をさせた．ここでは，日本語能力試験 1 級のテスト 70 項目への 197 人の受検者の反応データを用いる．具体的には，受検者 197 人中 97 人をランダム・サンプリングし，予測システムのアイテム・バンク用データとした．システム Ⓐ, Ⓑ, Ⓒ を用いて構成されたテストについて，バリデーション用に残した受検者 100 人のデータを用いて予測された平均得点とその標準偏差及び平均所要時間とその標準偏差と実測値との誤差 RMSE を計算し，テスト間の多重比較（WSD 検定）を行った．結果より，平均得点，得点の標準偏差，平均所要時間，所要時間の標準偏差について，5%で本システムと他システムに有意差があり，本システムの予測機能が有意に効果があったことを示している．以上より，本提案による予測機能が，テスト構成時における平均得点とその標準偏差及び平均所要時間とその標準偏差の予測に対して有意に効果的であることが示された．

5.2 アンケート分析

本システムの機能を評価するためにアンケート調査も行った．本システムを用いたテスト構成者群 10 名と本システムを用いていないテスト構成者群 20 名に分けてアンケート結果の差の検定を行った．ただし，紙面の都合上，結果は当日示すことにする．

本システムを用いることにより各予測作業が容易に行えるようになったと考えていることが有意水準 1% 以内でいえることを示している．また，全員への「本システムが有用であるか」というアンケートに対し，高い評点が与えられ，本システムの有効性を示している．

6. おわりに

本論では，e テスティングにおける得点・時間予測システムの提案を行った．具体的には，1) ベータ二項分布，混合二項分布を組み合わせて拡張した混合ベータ二項分布，2) 項目反応理論（ラッシュ・モデル，2 パラメータ・ロジスティック・モデル）を用いたテスト得点分布を提案し，従来に用いられてきた得点分布と比較を行った．テスト予測所要時間モデルとして，これまで提案されてきた所要時間分布による比較を行った．その結果，テスト得点予測モデルとして項目反応理論（2 パラメータ・ロジスティック・モデル），テスト所要時間予測モデルとして拡張ガンマ分布モデルが最も良い予測精度を示した．

以上の結果を用いて，テスト構成過程における構成されたテストの予測得点分布，予測所要時間分布の状態を逐次可視化するウェブ・ベース・テスト構成支援システムを開発した．評価実験の結果からシステムの有効性を示した．

参考文献

(1)永岡慶三，“予測機能を有する実用コンピュータ・テスト・システムの開発研究”，日本教育工学会論文誌，24(1)，2000，pp.63-72.

(2)M. Ueno，“Web-based computerized testing system for distance education”，Educational Technology Research，28，2005，pp.59-69.

(3)J.A. Keats and F.M. Lord,“A theoretical distribution for mental test scores”，Psychometrika，27，1962，pp.59‐72．

(4)永岡慶三，呉亜棟，“コンピュータ・テスティングにおける回答所要時間についての分析”，日本教育工学雑誌，12(4)，1989，pp.129-137.

(5)植野真臣，“ガンマ分布による e ラーニング所要時間データのオンライン解析”，日本教育工学会論文誌，29(2)，2005，pp.107-117.

(6)D. Thissen，“Timed testing: An approach using item response theory”，In D. Weiss (Ed.) New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing. N.Y.: Academic Press，1983, pp.179-203.

(7) N.D. Verhelst，H.H.F.M. Verstralen，and M.G.H. Jansen，“A logistic model for time-limit tests”，In W.J. van der Linden and R.K. Hambleton (Eds.)，Handbook of Modern Item Response Theory，New York: Springer，1997, pp.169-185．

(8) E.E. Roskam，“Models for speed and time-limit tests”，In W.J. van der Linden and R.K. Hambleton (Eds.) Handbook of Modern Item Response Theory，New York: Springer，1997, pp.187-208.

(9)P. Songmuang and M. Ueno，“e-Testing Management System”，Proc. of e-learn2005，2005, pp.3139-3148.

(10) G. Rasch,“An item analysis which takes individual differences into account”，British Journal of Mathematical and Statistical Psychology，19，1966, pp.49-57.

(11) F.M. Lord and M.R. Novick， Statistical Theories of Mental Test Scores，Addison-Wesley，Massachusetts ,1968.